# CRIME PREDICTION AND ANALYSIS USING MACHINE LEARNING

Submitted by

Mukilan.P
Antany Belgis
Shelly Niranjan
Mohamed Farook

**In the partial fulfillment of the award of the degree**

**of**

**BACHELOR OF TECHNOLOGY**

**in**

## ARTIFICIAL INTELLIGENCE AND DATA SCIENCE

## CARE COLLEGE OF ENGINEERING,TIRUCHIRAPALLI 6200009

## ANNA UNIVERSITY : CHENNAI 600 025

**NOVEMBER 2023**

# ABSTRACT

Crime analysis and prediction is a systematic approach for identifying the crime. This system can predict region which have high probability for crime occurrences and visualize crime prone area. Using the concept of data mining we can extract previously unknown, useful information from an unstructured data. The extraction of new information is predicted using the existing datasets. Crimes are treacherous and common social problem faced worldwide. Crimes affect the quality of life, economic growth and reputation of nation. With the aim of securing the society from crimes, there is a need for advanced systems and new approaches for improving the crime analytics for protecting their communities. We propose a system which can analysis, detect, and predict various crime probability in given region. This paper explains various types of criminal analysis and crime prediction using several data mining techniques.

# Table of contents


| Chapter No. | Title | Page No. |
|---|---|---|
| | **ABSTRACT** | **v** |
| | **LIST OF FIGURES** | **vii** |
| **1.** | **INTRODUCTION** | 1 |
| | 1.1 Overview | 1 |
| | 1.2 Scope of the Project | 2 |
| | 1.3 Objective of Project | 2 |
| | 1.4 Domain Overview | 2 |
| | 1.4.1 Machine Learning | 3 |
| | 1.4.2 Proposed Algorithm | 6 |
| **2.** | LITERATURE SURVEY | 14 |
| | 2.1 Analysis of the literature | 14 |
| | 2.2 Literary Reviews | 15 |
| **3.** | SYSTEM ANALYSIS | 19 |
| | 3.1 Existing System | 19 |
| | 3.2 Proposed System | 19 |
| | 3.3 System Architecture | 20 |
| | 3.4 Dataflow Diagram | 21 |
| | 3.5 System Requirement | 22 |
| | 3.6 Software Description | 22 |
| | 3.6.1 Application of Python | 23 |
| | 3.6.2 Features of Python | 23 |
| **4.** | SYSTEM IMPLEMENTATION | 24 |
| | 4.1 List of Modules | 24 |
| | 4.2 System Study | 25 |

| | | |
|---|---|---|
| **5.** | SOFTWARE TESTING | 27 |
| | 5.1 General | 28 |
| | 5.2 Test Driven Development | 29 |
| | 5.3 Unit Testing | 29 |
| | 5.4 Black Box Testing | 30 |
| | 5.5 Integration Testing | 30 |
| | 5.6 System Testing | 31 |
| | 5.7 Regression Testing | 32 |
| **6.** | CONCLUSION AND FUTURE WORK | 34 |
| | 6.1 Conclusion | 34 |
| | 6.2 Future Enhancement | 34 |
| | REFFERENCES | 35 |
| | APPENDIX | 36 |
| | A. Screenshots | 36 |
| | B. Sample Code | 43 |
| | C. Plagiarism report | 46 |

# LIST OF FIGURES


| FIGURE No. | FIGURE NAME | PAGE No. |
|---|---|---|
| 1.1 | Supervised Architecture | 4 |
| 1.2 | Block Diagram | 5 |
| 1.3 | Structure Of Decision Tree | 7 |
| 1.4 | Structure Of Random Forest | 10 |
| 1.5 | Random Forest Classifier | 11 |
| 1.6 | Logistic Regression Graph | 13 |
| 3.1 | Architecture Diagram | 20 |
| 3.2 | Dataflow Diagram | 21 |
| 4.1 | Dataset | 24 |
| 5.1 | Integration Testing | 31 |
| 5.2 | Module Testing | 31 |

# CHAPTER 1
# INTRODUCTION

## 1.1 OVERVIEW

Day by day crime data rate is increasing because the modern technologies and hi-tech methods are helps the criminals to achieving the illegal activities .according to Crime Record Bureau crimes like burglary, arson etc have been increased while crimes like murder, sex, abuse, gang rap etc have been increased. Crime data will be collected from various blogs, news and websites. The huge data is used as a record for creating a crime report database. The knowledge which is acquired from the data mining techniques will help in reducing crimes as it helps in finding the culprits faster and also the areas that are most affected by crime .Data mining helps in solving the crimes faster and this technique gives good results when applied on crime dataset, the information obtained from the data mining techniques can help the police department. A particular approach has been found to be useful by the police, which is the identification of crime ‘hot spots ‘which indicates areas with a high concentration of crime. Use of data mining techniques can produce important results from crime report datasets. The very step in study of crime is crime analysis. Crime analysis is exploring, inter relating and detecting relationship between the various crimes and characteristics f the crime. This analysis helps in preparing statistics, queries and maps on demand. It also helps to see if a crime in a certain known pattern or a new pattern necessary. Crimes can be predicted as the criminal are active and operate in their comfort zones. Once successful they try to replicate the crime under similar circumstances. The occurrences of crime depended on several factors such as intelligence of criminals, security of a location, etc The work has followed the steps that used in data analysis, in which the important phases are Data collection, data classification, pattern identification, prediction and visualization. The proposed framework uses different visualization techniques to show the trends of crimes and various ways that can predicts the crime using machine learning algorithm. The inputs to our algorithms are time (hour, day, month, and year), place (latitude and longitude), and class of crime:

- Act 379 - Robbery
- Act 13 – Gambling
- Act 279 - Accident
- Act 323 - Violence
- Act 302 - Murder
- Act 363 - Kidnapping

The output is the class of crime that is likely to have occurred. We try out multiple classification algorithms, such as KNN (K-Nearest Neighbors), Decision Trees, and Random Forests. We also perform multiple classification tasks – we first try to predict which of 6 classes of crimes are likely to have occurred, and later try to differentiate between violent and non-violent crimes.

## 1.2 SCOPE OF THE PROJECT

Much of the current work is focused in two major directions:

- Predicting surges and hotspots of crime, and
- Understanding patterns of criminal behavior that could help in solving criminal investigations.

## 1.3 OBJECTIVE OF PROJECT

The objective of our work is to:

- Predicting crime before it takes place.
- Predicting hotspots of crime.
- Understanding crime pattern.
- Classify crime based on location.
- Analysis of crime in Indore

## 1.4 DOMAIN OVERVIEW

Machine Learning is the most popular technique of predicting the future or classifying information to help people in making necessary decisions. Machine Learning algorithms are trained over instances or examples through which they learn from past experiences and also analyze the historical data. Therefore, as it trains over the examples, again and again, it is able to identify patterns in order to make predictions about the future. Data is the core backbone of machine learning algorithms. With the

concept of Machine Learning that learns from the historical images through which they are capable of generating more images. This is also applied towards speech and text synthesis. Therefore, Machine Learning has opened up a vast potential for data science applications.

### 1.4.1 MACHINE LEARNING

Machine Learning combines computer science, mathematics, and statistics. Statistics is essential for drawing inferences from the data. Mathematics is useful for developing machine learning models and finally, computer science is used for implementing algorithms. However, simply building models is not enough. You must also optimize and tune the model appropriately so that it provides you with accurate results. Optimization techniques involve tuning the hyper parameters to reach an optimum result.

The world today is evolving and so are the needs and requirements of people. Furthermore, we are witnessing a fourth industrial revolution of data. In order to derive meaningful insights from this data and learn from the way in which people and the system interface with the data, we need computational algorithms that can churn the data and provide us with results that would benefit us in various ways. Machine Learning has revolutionized industries like medicine, healthcare, manufacturing, banking, and several other industries. Therefore, Machine Learning has become an essential part of modern industry. Data is expanding exponentially and in order to harness the power of this data, added by the massive increase in computation power, Machine Learning has added another dimension to the way we perceive information. Machine Learning is being utilized everywhere. The electronic devices you use, the applications that are part of your everyday life are powered by powerful machine learning algorithms.

With an exponential increase in data, there is a need for having a system that can handle this massive load of data. Machine Learning models like Deep Learning allow the vast majority of data to be handled with an accurate generation of predictions. Machine Learning has revolutionized the way we perceive information and the various insights we can gain out of it. These machine learning algorithms use the patterns contained in the training data to perform classification and future predictions.

Whenever any new input is introduced to the ML model, it applies its learned patterns over the new data to make future predictions. Based on the final accuracy, one

Types of Machine Learning

Machine Learning Algorithms can be classified into 3 types as follows –

- Supervised Learning
- Unsupervised Learning
- Reinforcement Learning

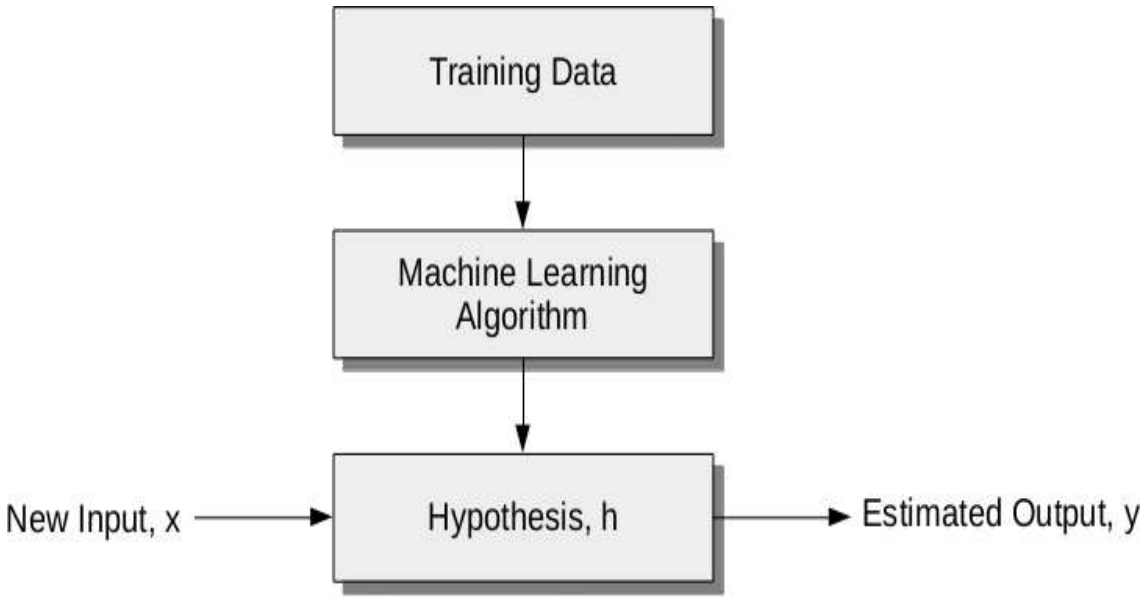



**Fig -1.1 Supervised Architecture**

**SUPERVISED LEARNING**

In the majority of supervised learning applications, the ultimate goal is to develop a finely tuned predictor function h(x) (sometimes called the "hypothesis"). "Learning" consists of using sophisticated mathematical algorithms to optimize this function so that, given input data x about a certain domain (say, square footage of a house), it will accurately predict some interesting value h(x) (say, market price for said house).

$$h(x_1, x_2, x_3, x_4) = \theta_0 + \theta_1 x_1 + \theta_2 x_3^2 + \theta_3 x_3 x_4 + \theta_4 x_1^3 x_2^2 + \theta_5 x_2 x_3^4 x_4^2$$

This function takes input in four dimensions and has a variety of polynomial terms. Deriving a normal equation for this function is a significant challenge. Many modern machine learning problems take thousands or even millions of dimensions of data to build predictions using hundreds of coefficients. Predicting how an organism's genome will be expressed, or what the climate will be like in fifty years, are examples of such complex problems.

Under supervised ML, two major subcategories are:

- Regression machine learning systems: Systems where the value being predicted falls somewhere on a continuous spectrum.
- Classification machine learning systems: Systems where we seek a yes-or-no prediction.

In practice, x almost always represents multiple data points. So, for example, a housing price predictor might take not only square-footage (x1) but also number of bedrooms (x2), number of bathrooms (x3), number of floors (x4), year built (x5), zip code (x6), and so forth. Determining which inputs to use is an important part of ML design. However, for the sake of explanation, it is easiest to assume a single input value is used.

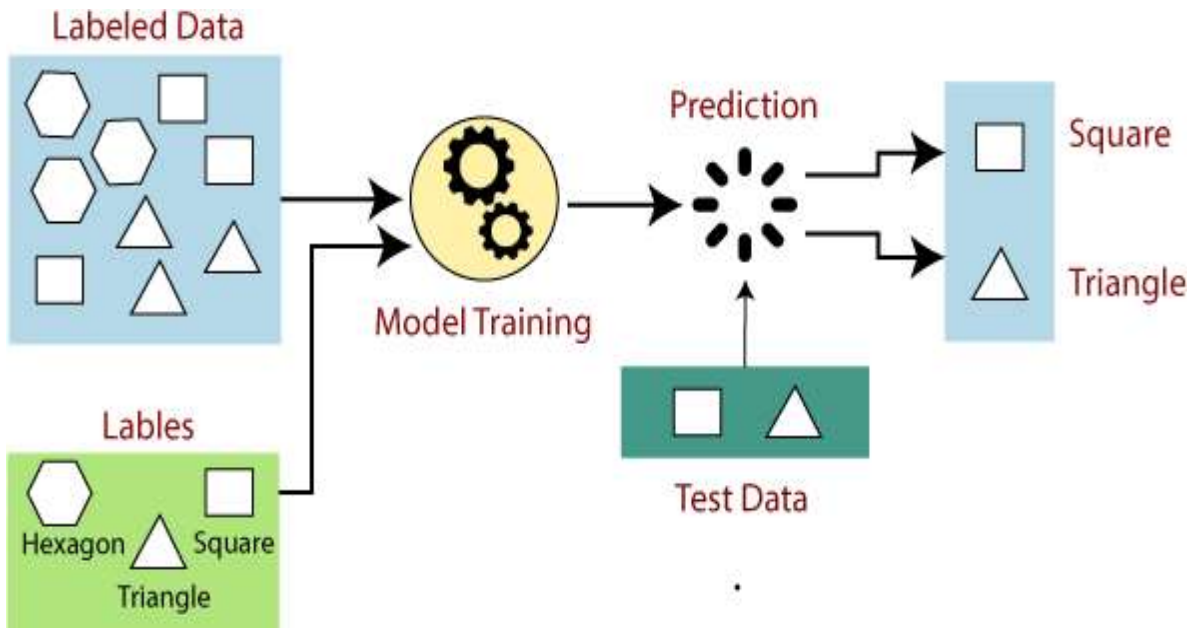



**Fig-1.2 Block Diagram**

Steps Involved in Supervised Learning:

- First Determine the type of training dataset
- Collect/Gather the labelled training data.
- Split the training dataset into training dataset, test dataset, and validation dataset.
- Determine the input features of the training dataset, which should have enough knowledge so that the model can accurately predict the output.
- Determine the suitable algorithm for the model, such as support vector machine, decision tree, etc.
- Execute the algorithm on the training dataset. Sometimes we need validation sets as the control parameters, which are the subset of training datasets.

**REGRESSION**

Regression algorithms are used if there is a relationship between the input variable and the output variable. It is used for the prediction of continuous variables, such as Weather forecasting, Market Trends, etc.

- Linear Regression
- Regression Trees
- Non-Linear Regression
- Bayesian Linear Regression
- Polynomial Regression

**CLASSIFICATION**

Classification algorithms are used when the output variable is categorical, which means there are two classes such as Yes-No, Male-Female, True-false, etc.

Spam Filtering,

- Random Forest
- Decision Tree
- Logistic Regression
- Support vector Machines

### 1.4.2 PROPOSED ALGORITHMS

**Decision Tree Classification Algorithm**

- Decision Tree is a supervised learning technique that can be used for both classification and Regression problems, but mostly it is preferred for solving Classification problems. It is a tree-structured classifier, where internal nodes represent the features of a dataset, branches represent the decision rules and each leaf node represents the outcome.
- In a Decision tree, there are two nodes, which are the Decision Node and Leaf Node. Decision nodes are used to make any decision and have multiple branches, whereas Leaf nodes are the output of those decisions and do not contain any further branches.
- The decisions or the test are performed on the basis of features of the given dataset.
- It is a graphical representation for getting all the possible solutions to a

node, which expands on further branches and constructs a tree-like structure.

- In order to build a tree, we use the CART algorithm, which stands for Classification and Regression Tree algorithm.
- A decision tree simply asks a question, and based on the answer (Yes/No), it further split the tree into sub trees.
- Below diagram explains the general structure of a decision tree:

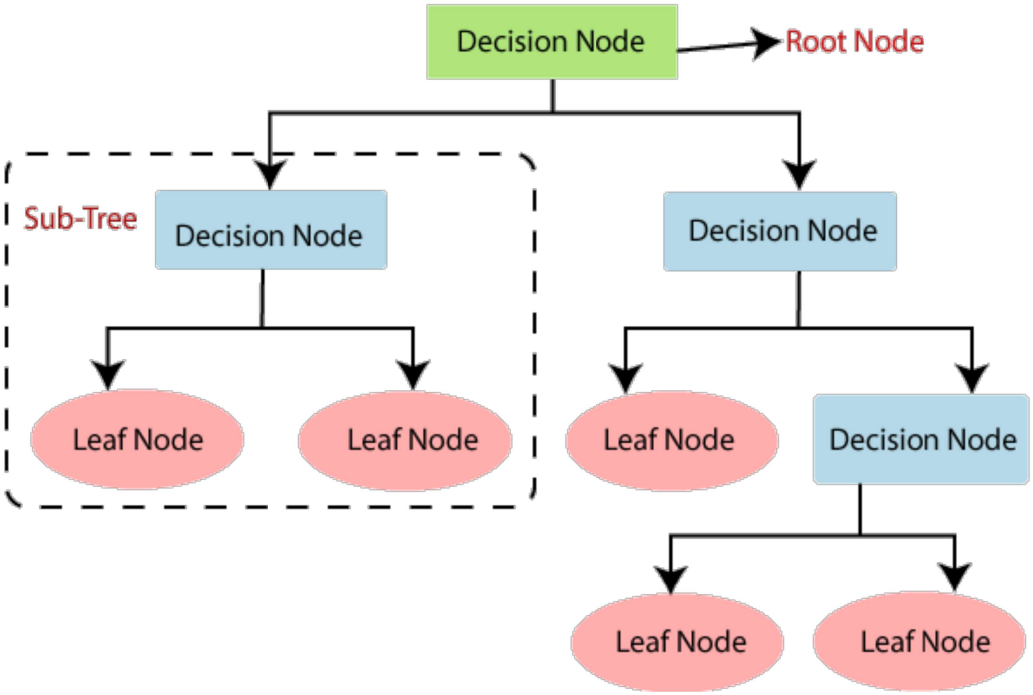



**Fig-1.3 Structure of decision tree**

There are various algorithms in Machine learning, so choosing the best algorithm for the given dataset and problem is the main point to remember while creating a machine learning model. Below are the two reasons for using the Decision tree:

- Decision Trees usually mimic human thinking ability while making a decision, so it is easy to understand.
- The logic behind the decision tree can be easily understood because it shows a tree-like structure.

**Decision Tree Terminologies**

- Root node is from where the decision tree starts. It represents the entire dataset, which further gets divided into two or more homogeneous sets.
- Leaf nodes are the final output node, and the tree cannot be segregated further after getting a leaf node.

according to the given conditions.

- A tree formed by splitting the tree known as branch tree.
- Pruning is the process of removing the unwanted branches from the tree.
- The root node of the tree is called the parent node, and other nodes are called the child nodes.

In a decision tree, for predicting the class of the given dataset, the algorithm starts from the root node of the tree. This algorithm compares the values of root attribute with the record (real dataset) attribute and, based on the comparison, follows the branch and jumps to the next node.

For the next node, the algorithm again compares the attribute value with the other sub-nodes and move further. It continues the process until it reaches the leaf node of the tree.

**Python Implementation of Decision Tree**

Now we will implement the Decision tree using Python. For this, we will use the dataset "user_data.csv," which we have used in previous classification models. By using the same dataset, we can compare the Decision tree classifier with other classification models such as KNN, SVM, Logistic Regression, etc.

Steps will also remain the same, which are given below:

- Data Pre-processing step
- Fitting a Decision-Tree algorithm to the Training set
- Predicting the test result
- Test accuracy of the result(Creation of Confusion matrix)
- Visualizing the test set result.

**Advantages of the Decision Tree**

- It is simple to understand as it follows the same process which a human follow while making any decision in real-life.
- It can be very useful for solving decision-related problems.
- It helps to think about all the possible outcomes for a problem.
- There is less requirement of data cleaning compared to other algorithms.

**Disadvantages of the Decision Tree**

- The decision tree contains lots of layers, which makes it complex.
- It may have an over fitting issue, which can be resolved using the Random Forest algorithm.
- For more class labels, the computational complexity of the decision tree may increase.

**RANDOM FOREST ALGORITHM**

Random forest algorithm can use both for classification and the regression kind of problems. In this you are going to learn, how the random forest algorithm works in machine learning for the classification task.

Random Forest is a popular machine learning algorithm that belongs to the supervised learning technique. It can be used for both Classification and Regression problems in ML. It is based on the concept of ensemble learning, which is a process of combining multiple classifiers to solve a complex problem and to improve the performance of the model.

A random forest algorithm consists of many decision trees. The 'forest' generated by the random forest algorithm is trained through bagging or bootstrap aggregating. Bagging is an ensemble meta-algorithm that improves the accuracy of machine learning algorithms.

As the name suggests, "Random Forest is a classifier that contains a number of decision trees on various subsets of the given dataset and takes the average to improve the predictive accuracy of that dataset." Instead of relying on one decision tree, the random forest takes the prediction from each tree and based on the majority votes of predictions, and it predicts the final output.

The below diagram explains the working of the Random Forest algorithm:

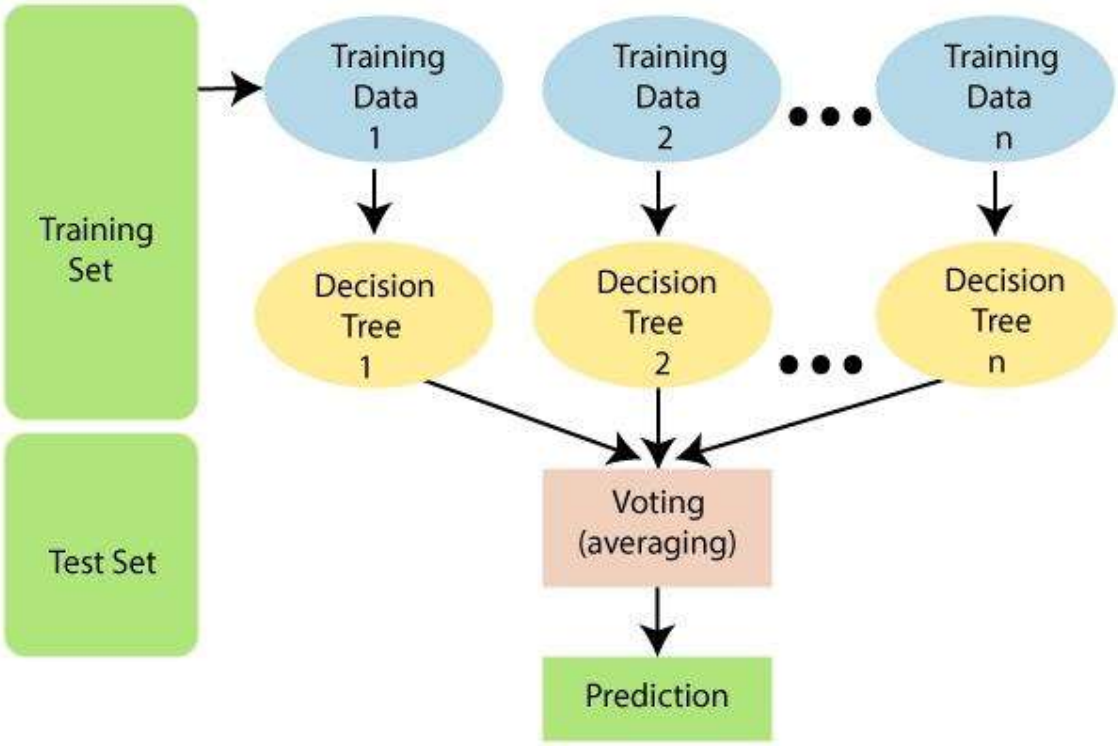



**Fig-1.4 Structure of Random Forest**

Below are some points that explain why we should use the Random Forest algorithm:

- It takes less training time as compared to other algorithms.
- It predicts output with high accuracy, even for the large dataset it runs efficiently.
- It can also maintain accuracy when a large proportion of data is missing.

**Features of a Random Forest Algorithm**

- It's more accurate than the decision tree algorithm.
- It provides an effective way of handling missing data.
- It can produce a reasonable prediction without hyper-parameter tuning.
- It solves the issue of over fitting in decision trees.
- In every random forest tree, a subset of features is selected randomly at the node's splitting point.

**Classification in random forests**

Classification in random forests employs an ensemble methodology to attain the outcome. The training data is fed to train various decision trees. This dataset consists of observations and features that will be selected randomly during the splitting of nodes.

A rain forest system relies on various decision trees. Every decision tree consists of

the majority-voting system. In this case, the output chosen by the majority of the decision trees becomes the final output of the rain forest system. The diagram below shows a simple random forest classifier.

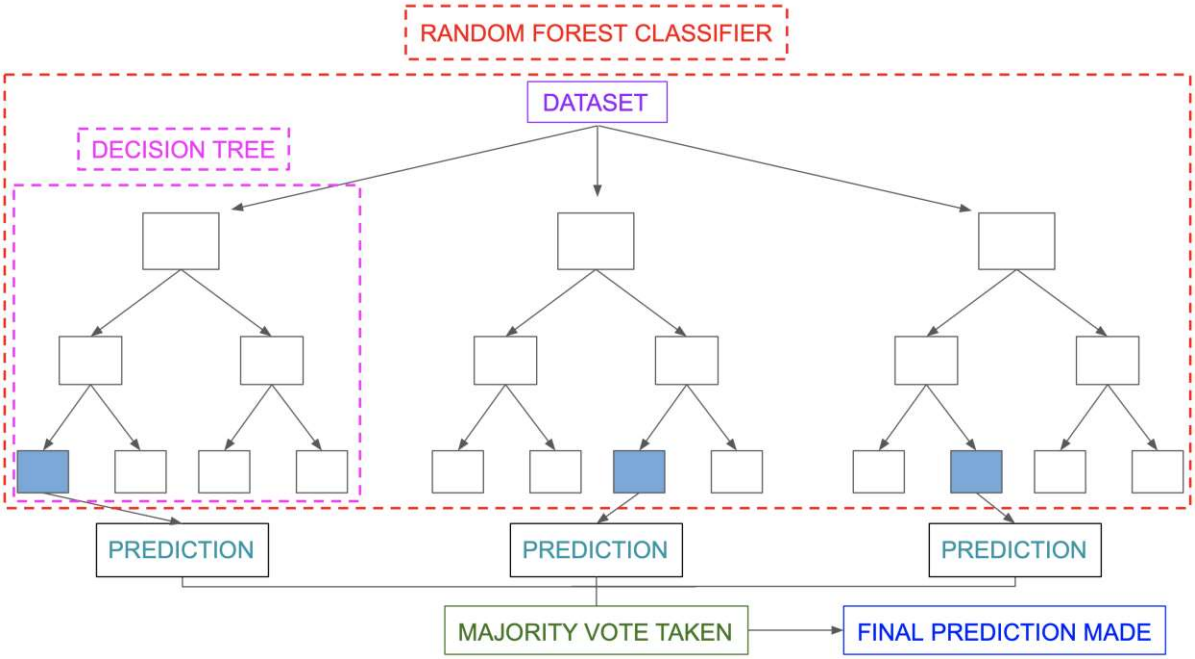



**Fig-1.5 Random Forest Classifier**

**Random Forest Steps**

- Randomly select “k” features from total “m” features,where k << m
- Among the “k” features, calculate the node “d” using the best split point.
- Split the node into daughter nodes using the best split.
- Repeat 1 to 3 steps until “l” number of nodes has been reached.
- Build forest by repeating steps 1 to 4 for “n” number times to create “n” number of trees.

The beginning of random forest algorithm starts with randomly selecting “k” features out of total “m” features. In the image, you can observe that we are randomly taking features and observations.

**Applications of Random Forest**

There are mainly four sectors where Random forest mostly used:

- **Banking:** Banking sector mostly uses this algorithm for identification of loan risk.
- **Medicine:** With the help of this algorithm, disease trends and risks of the disease can be identified.
- **Land Use:** We can identify the areas of similar land use by this algorithm.
- **Marketing:** Marketing trends can be identified using this algorithm.

**Advantages of Random Forest**

- Random Forest is capable of performing both Classification and Regression tasks.
- It is capable of handling large datasets with high dimensionality.
- It enhances the accuracy of the model and prevents the over fitting issue.

**Disadvantages of Random Forest**

- Although random forest can be used for both classification and regression tasks, it is not more suitable for Regression tasks.

**LOGISTIC REGRESSION**

- Logistic regression is one of the most popular Machine Learning algorithms, which comes under the Supervised Learning technique. It is used for predicting the categorical dependent variable using a given set of independent variables.
- Logistic regression predicts the output of a categorical dependent variable. Therefore the outcome must be a categorical or discrete value. It can be either Yes or No, 0 or 1, true or False, etc. but instead of giving the exact value as 0 and 1, it gives the probabilistic values which lie between 0 and 1.
- Logistic Regression is much similar to the Linear Regression except that how they are used. Linear Regression is used for solving Regression problems, whereas Logistic regression is used for solving the classification problems.
- In Logistic regression, instead of fitting a regression line, we fit an "S" shaped

- The curve from the logistic function indicates the likelihood of something such as whether the cells are cancerous or not, a mouse is obese or not based on its weight, etc.
- Logistic Regression is a significant machine learning algorithm because it has the ability to provide probabilities and classify new data using continuous and discrete datasets.
- Logistic Regression can be used to classify the observations using different types of data and can easily determine the most effective variables used for the classification.

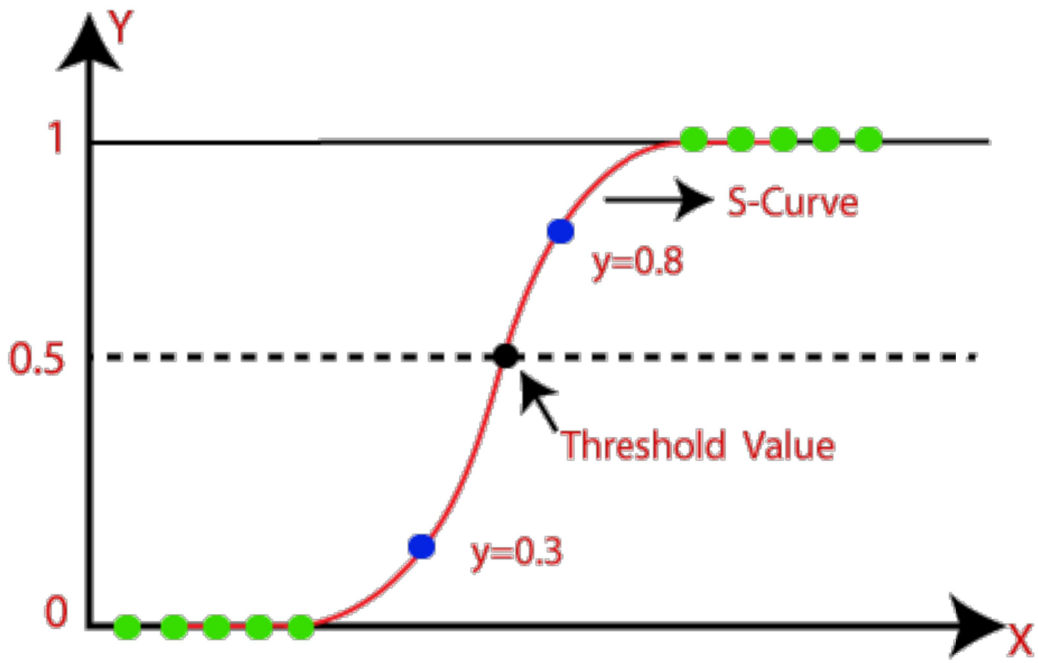



**Fig-1.6 Logistic Regression graph**

# CHAPTER 2

# LITERATURE SURVEY

## 2.1 ANALYSIS OF THE LITERATURE

Literature survey is the main advance in programming improvement measure. Prior to building up the instrument it is important to decide the time factor, economy and friends strength. When these things are fulfilled, at that point the subsequent stage is to figure out which working framework and language can be utilized for building up the device. When the developers begin assembling the apparatus the software engineers need parcel of outer help. This help can be gotten from senior developers, from book or from sites. The major part of the project development sector considers and fully survey all the required needs for developing the project. Before developing the tools and the associated designing it is necessary to determine and survey the time factor, resource requirement, man power, economy, and company strength. Prior to building the framework the above thought are considered for building up the proposed framework. The significant piece of the undertaking advancement area considers and completely survey all the necessary requirements for building up the venture. For each undertaking Literature survey is the main area in programming improvement measure. Prior to building up the instruments and the related planning it is important to decide and survey the time factor, asset prerequisite, labor, economy, and friends strength. When these things are fulfilled and completely surveyed, at that point the following stage is to decide about the product details in the separate framework, for example, what kind of working framework the venture would require and what are largely the important programming are expected to continue with the subsequent stage like building up the apparatuses, and the related activities. Here we have taken the general surveys of different creators and noted down the fundamental central issues with respect to their work. In this venture literature survey assumes a prevailing part in get assets from different areas and all the connected points that are exceptionally valuable under this segment. The most awesome aspect if this is the manner in which things get all together and encourages us to suite our work according to the current information.

## 2.2 LITERARY REVIEWS

***"An Exploration of Crime Prediction Using Data Mining on Open Data"Ginger Saltos and Minhaela Cocea (2017)***

The increase in crime data recording coupled with data analytics resulted in the growth of research approaches aimed at extracting knowledge from crime records to better understand criminal behavior and ultimately prevent future crimes. While many of these approaches make use of clustering and association rule mining techniques, there are fewer approaches focusing on predictive models of crime. In this paper, we explore models for predicting the frequency of several types of crimes by LSOA code (Lower Layer Super Output Areas — an administrative system of areas used by the UK police) and the frequency of anti-social behavior crimes. Three algorithms are used from different categories of approaches: instance-based learning, regression and decision trees. The data are from the UK police and contain over 600,000 records before preprocessing. The results, looking at predictive performance as well as processing time, indicate that decision trees (M5P algorithm) can be used to reliably predict crime frequency in general as well as anti-social behavior frequency.The experiments were conducted using the SCIAMA High Performance Computer Cluster at the University of Portsmouth.

***"Crime Analysis and Prediction Using Data Mining" Shiju Sathyadevan, Devan M.S, Surya Gangadharan (IEEE-2014)***

Crime analysis and prevention is a systematic approach for identifying and analyzing patterns and trends in crime. Our system can predict regions which have high probability for crime occurrence and can visualize crime prone areas. With the increasing advent of computerized systems, crime data analysts can help the Law enforcement officers to speed up the process of solving crimes. Using the concept of data mining we can extract previously unknown, useful information from an unstructured data. Here we have approach between computer science and criminal justice to develop a data mining procedure that can help solve crimes faster. Instead of focusing on causes of crime occurrence like criminal background of offender, political enmity etc we are focusing mainly on crime factors of each day. This paper has tested the accuracy of classification and prediction based on different test sets. Classification is done based on the Bayes theorem which showed more than 90% accuracy.

***"Crime Pattern Analysis, Visualization And Prediction Using Data Mining" Rajkumar.S, Sakkarai , Soundarya Jagan.J, Varnikasree.P (2015)***

Crime against women these days has become problem of every nation around the globe many countries are trying to curb this problem. Preventive are taken to reduce the increasing number of cases of crime against women. A huge amount of data set is generated every year on the basis of reporting of crime. This data can prove very useful in analyzing and predicting crime and help us prevent the crime to some extent. Crime analysis is an area of vital importance in police department. Study of crime data can help us analyze crime pattern, inter-related clues& important hidden relations between the crimes. That is why data mining can be great aid to analyze, visualize and predict crime using crime data set. Classification and correlation of data set makes it easy to understand similarities & dissimilarities amongst the data objects. We group data objects using clustering technique. Dataset is classified on the basis of some predefined condition. Here grouping is done according to various types of crimes against women taking place in different states and cities of India. Crime mapping will help the administration to plan strategies for prevention of crime, further using data mining technique data can be predicted and visualized in various form in order to provide better understanding of crime patterns.

***"Survey paper on crime prediction using ensemble approach" Ayisheshim Almaw,Kalyani kadam (2018)***

Crime is a foremost problem where the top priority has been concerned by individual, the community and government. This paper investigates a number of data mining algorithms and ensemble learning which are applied on crime data mining. This survey paper describes a summary of the methods and techniques which are implemented in crime data analysis and prediction. Crime forecasting is a way of trying to mining out and decreasing the upcoming crimes by forecasting the future crime that will occur. Crime prediction practices historical data and after examining data, predict the upcoming crime with respect to location, time, day, season and year. In present crime cases rapidly increases so it is an inspiring task to foresee upcoming crimes closely with better accuracy. Data mining methods are too important to resolving crime problem with investigating hidden crime patterns.so the objective of this study could be analyzing and discussing various methods which are applied on crime prediction and analysis. This paper delivers reasonable investigation of Data mining Techniques and ensemble

***"Survey on crime analysis and predict ion using data mining techniques" Benjamin Fredrick David. H and Suruliand I (2017)***

Data Mining is the procedure which includes evaluating and examining large pre-existing databases in order to generate new information which may be essential to the organization. The extraction of new information is predicted using the existing datasets. Many approaches for analysis and prediction in data mining had been performed. But, many few efforts has made in the criminology field. Many few have taken efforts for comparing the information all these approaches produce. The police stations and other similar criminal justice agencies hold many large databases of information which can be used to predict or analyze the criminal movements and criminal activity involvement in the society. The criminals can also be predicted based on the crime data. The main aim of this work is to perform a survey on the supervised learning and unsupervised learning techniques that has been applied towards criminal identification. This paper presents the survey on the Crime analysis and crime prediction using several Data Mining techniques. The quantitative analysis produced results which shows the increase in the Accuracy level of classification because of using the GA to optimize the parameters.

***"Systematic Literature Review of Crime Prediction and Data Mining" Falade Adesola and Ambrose Azeta (2019)***

Using crime datasets requires different strategies for the varying types of data that describe illicit activity. Falade et al. (2019) provide a survey of crime prediction efforts wherein various machine learning methods have been applied to multiple types of datasets: criminal records, social media, news, and police reports. The authors note the different opportunities and challenges that each type of crime dataset presents, such as social media posts being highly unstructured and First Information Reports (FIRs) being unstructured but reliable. This paper is explains techniques used, challenges addressed, methodologies used, and crime data mining and analysis paper. The methodologies is composed of three stages the first stage involves the research work related to crime data mining, second stage is concerned with establishing a classification and the third stage is involves the presentation of summary of research in crime data mining and analysis and report of this survey.

***"Crime Detection Techniques Using data Mining and K-Means" Khushabu A.Bokde, Tisksha P. Kakade, Dnyanes hwari S. Tumsare, Chetan G. Wadhai(2018)***

Crimes will somehow influence organizations and institutions when occurred frequently in a society. Thus, it seems necessary to study reasons, factors and relations between occurrence of different crimes and finding the most appropriate ways to control and avoid more crimes. The main objective of this paper is to classify clustered crimes based on occurrence frequency during different years. Data mining is used extensively in terms of analysis, investigation and discovery of patterns for occurrence of different crimes. We applied a theoretical model based on data mining techniques such as clustering and classification to real crime dataset recorded by police in England and Wales within 1990 to 2011. We assigned weights to the features in order to improve the quality of the model and remove low value of them. The Genetic Algorithm (GA) is used for optimizing of Outlier Detection operator parameters using Rapid Miner tool.

***"Empirical Analysis for Crime Prediction and Forecasting Using Machine Learning and Deep Learning Techniques" Wajiha Safat, Sohail Asghar, Saira Andleeb Gillani (IEEE-2021)***

Crime and violation are the threat to justice and meant to be controlled. Accurate crime prediction and future forecasting trends can assist to enhance metropolitan safety computationally. The limited ability of humans to process complex information from big data hinders the early and accurate prediction and forecasting of crime. The accurate estimation of the crime rate, types and hot spots from past patterns creates many computational challenges and opportunities. Despite considerable research efforts, yet there is a need to have a better predictive algorithm, which direct police patrols toward criminal activities. Previous studies are lacking to achieve crime forecasting and prediction accuracy based on learning models. Therefore, this study applied different machine learning algorithms, namely, the logistic regression, support vector machine (SVM), Naïve Bayes, k-nearest neighbors (KNN), decision tree, multilayer perceptron (MLP), random forest, and extreme Gradient Boosting, and time series analysis by long-short term memory (LSTM) and autoregressive integrated moving average (ARIMA) model to better fit the crime data. The performance of LSTM for time series analysis was reasonably adequate in order of magnitude of root mean square error (RMSE) and mean absolute error (MAE), on both data sets. Exploratory data analysis predicts more than 35 crime types and overall, these results provide early identification of crime, hot

# CHAPTER 3
# SYSTEM ANALYSIS

## 3.1 EXISTING SYSTEM

Data mining in the study and analysis of criminology can be categorized into main areas, crime control and crime suppression. De Bruin et. Al. introduced a framework for crime trends using a new distance measure for comparing all individuals based on their profiles and then clustering them accordingly. Manish Gupta et. Al. highlights the existing systems used by Indian police as e-governance initiatives and also proposes an interactive query based interface as crime analysis tool to assist police in their activities. He proposed interface which is used to extract useful information from the vast crime database maintained by National Crime Record Bureau (NCRB) and find crime hot spots using crime data mining techniques such as clustering etc. The effectiveness of the proposed interface has been illustrated on Indian crime records. Sutapat Thiprungsri examines the application of cluster analysis in the accounting domain, particularly discrepancy detection in audit. The purpose of his study is to examine the use of clustering technology to automate fraud filtering during an audit. He used cluster analysis to help auditors focus their efforts when evaluating group life insurance claims.

## 3.2 PROPOSED SYSTEM

In this project, we will be using the technique of machine learning and data science for crime prediction of crime data set. The crime data is extracted from the official portal of police. It consists of crime information like location description, type of crime, date, time, latitude, longitude. Before training of the model data preprocessing will be done following this feature selection and scaling will be done so that accuracy obtain will be high. The Logistic Regression classification and various other algorithms (Decision Tree and Random Forest) will be tested for crime prediction and one with better accuracy will be used for training. Visualization of dataset will be done in terms of graphical representation of many cases for example at which time the criminal rates are high or at which month the criminal activities are high. The whole purpose of this project is to give a just idea of how machine learning can be used by the law enforcement agencies to detect, predict and solve crimes at a much faster rate and thus reduces the crime rate. This can be used in other states or countries depending upon the availability of the dataset.

### 3.3 SYSTEM ARCHITECTURE

There are many kinds of architecture diagrams, like a software architecture diagram, system architecture diagram, application architecture diagram, security architecture diagram, etc.

For system developers, they need system architecture diagrams to understand, clarify, and communicate ideas about the system structure and the user requirements that the system must support.

It describes the overall features of the software is concerned with defining the requirements and establishing the high level of the system. During architectural design, the various web pages and their interconnections are identified and designed. The major software components are identified and decomposed into processing modules and conceptual data structures and the interconnections among the modules are identified. The following modules are identified in the proposed system.

The system architectural design is the design process for identifying the subsystems making up the system and framework for subsystem control and communication. The goal of the architectural design is to establish the overall structure of software system.

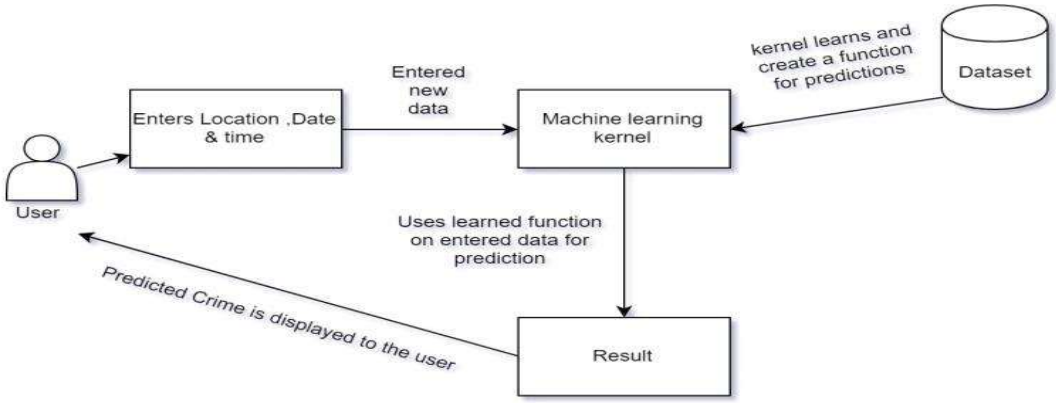



**Fig- 3.1 Architecture diagram**

### 3.4 DATA FLOW DIAGRAM:

- The DFD is also called as bubble chart. It is a simple graphical formalism that can be used to represent a system in terms of input data to the system, various processing carried out on this data, and the output data is generated by this system.
- The data flow diagram (DFD) is one of the most important modeling tools. It is used to model the system components. These components are the system process, the data used by the process, an external entity that interacts with the system and the information flows in the system.
- DFD shows how the information moves through the system and how it is modified by a series of transformations. It is a graphical technique that depicts information flow and the transformations that are applied as data moves from input to output.
- DFD is also known as bubble chart. A DFD may be used to represent a system at any level of abstraction. DFD may be partitioned into levels that represent increasing information flow and functional detail.

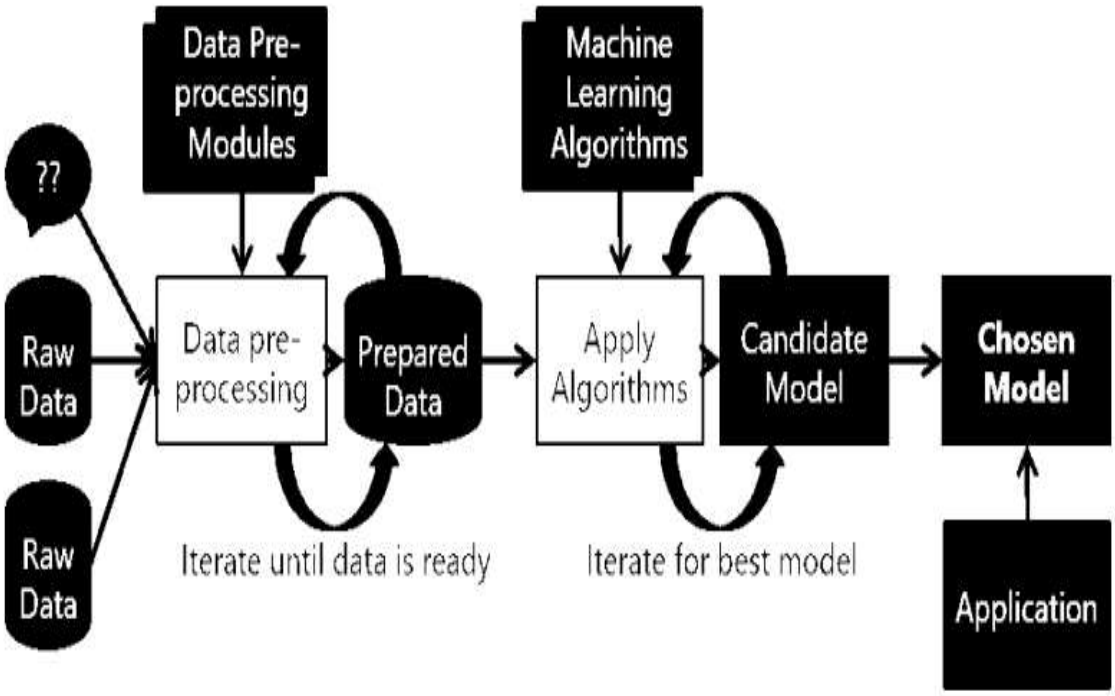



**Fig-3.2 Dataflow Diagram**

## 3.5 SYSTEM REQUIREMENTS

### HARDWARE REQUIREMENTS

System : intel Core i5.

Hard Disk : 1 TB.

Monitor : 15'' LED

Input Devices : Keyboard, Mouse

Ram : 8 GB.

### SOFTWARE REQUIREMENTS

Operating system : Windows 10.

Coding Language : Python

## 3.6 SOFTWARE DESCRIPTION

Python is a free, open-source programming language. Therefore, all you have to do is install Python once, and you can start working with it. Not to mention that you can contribute own code to the community. Python is also a cross-platform compatible language. So, what does this mean? Well, you can install and run Python on several operating systems. Whether you have a Windows, Mac or Linux, you can rest assure that Python will work on all these operating systems. Python is also a great visualization tool. It provides libraries such as Matplotlib, seaborn and bokeh to create stunning visualizations.

In addition, Python is the most popular language for machine learning and deep learning. As a matter of fact, today, all top organizations are investing in Python to implement machine learning in the back-end.

Python is a general-purpose interpreted, interactive, object-oriented, and high-level programming language. It was created by Guido van Rossum during 1985- 1990. Like Perl, Python source code is also available under the GNU General Public License (GPL).It was developed by Guido van Rossum in the late eighties and early nineties at the National Research Institute for Mathematics and Computer Science in the Netherlands.

Python is derived from many other languages, including ABC, Modula-3, C, C++, Algol-68, SmallTalk, and Unix shell and other scripting languages.It is copyrighted. Like Perl, Python source code is now available under the GNU General Public License (GPL.) It is now maintained by a core development team at the institute, although Guido

### 3.6.1 APPLICATIONS OF PYTHON

- **Easy-to-learn** − Python has few keywords, simple structure, and a clearly defined syntax. This allows the student to pick up the language quickly.
- **Easy-to-read** − Python code is more clearly defined and visible to the eyes.
- **Easy-to-maintain** − Python's source code is fairly easy-to-maintain.
- **A broad standard library** − Python's bulk of the library is very portable and cross-platform compatible on UNIX, Windows, and Macintosh.
- **Interactive Mode** − Python has support for an interactive mode which allows interactive testing and debugging of snippets of code.
- **Portable** − Python can run on a wide variety of hardware platforms and has the same interface on all platforms.
- **Extendable** − You can add low-level modules to the Python interpreter. These modules enable programmers to add to or customize their tools to be more efficient.
- **Databases** − Python provides interfaces to all major commercial databases.
- **GUI Programming** − Python supports GUI applications that can be created and ported to many system calls, libraries and windows systems, such as Windows MFC, Macintosh, and the X Window system of Unix.
- **Scalable** − Python provides a better structure and support for large programs than shell scripting.

### 3.6.2 FEATURES OF PYTHON

- It supports functional and structured programming methods as well as OOP.
- It can be used as a scripting language or can be compiled to byte-code for building large applications.
- It provides very high-level dynamic data types and supports dynamic type checking
- It supports automatic garbage collection.
- It can be easily integrated with C, C++, COM, ActiveX, CORBA, and Java.

# CHAPTER 4

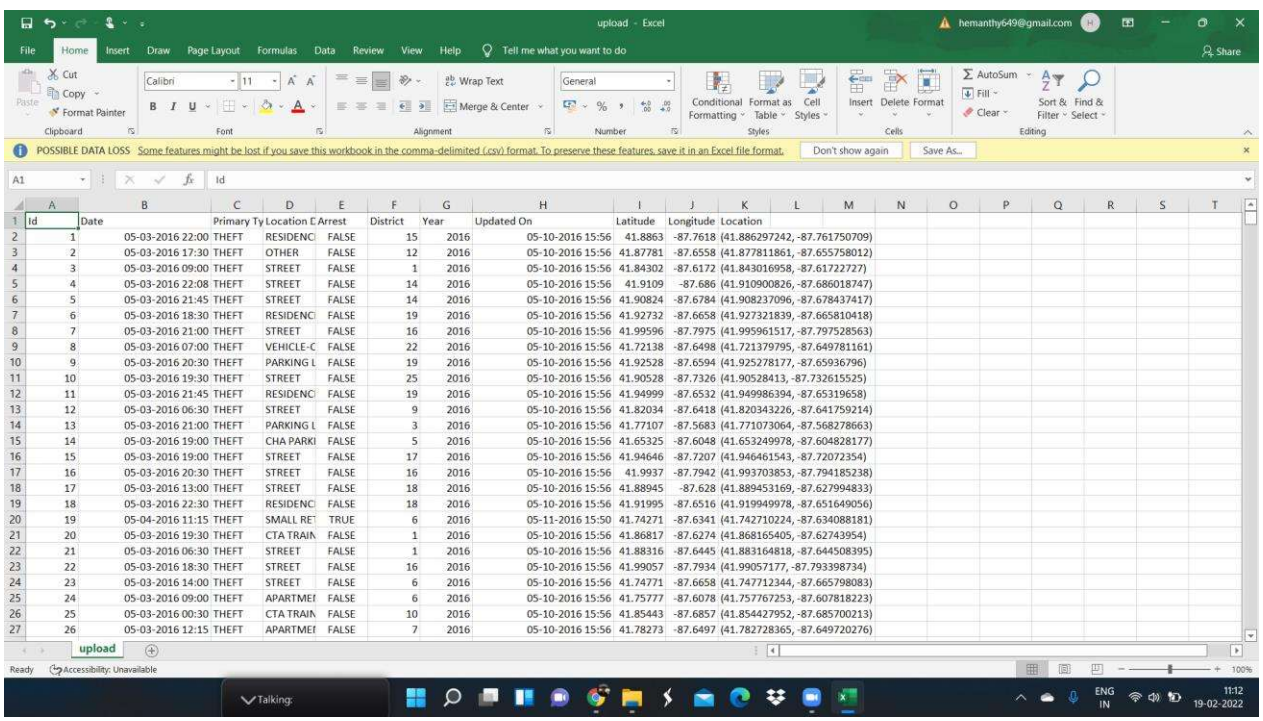
# SYSTEM IMPLEMENTATION

## 4.1 LIST OF MODULES

- Data Collection Module
- Data Preprocessing Module
- Feature selection Module
- Building and Training Model
- Prediction  Module
- Visualization Module

### Data collection Module

Crime dataset from kaggle having 8000 entries of crime data is used in CSV format.



| Id | Date | Primary Ty | Location D | Arrest | District | Year | Updated On | Latitude | Longitude | Location |
|---|---|---|---|---|---|---|---|---|---|---|
| 1 | 05-03-2016 22:00 | THEFT | RESIDENC | FALSE | 15 | 2016 | 05-10-2016 15:56 | 41.8863 | -87.7618 | (41.886297242, -87.761750709) |
| 2 | 05-03-2016 17:30 | THEFT | OTHER | FALSE | 12 | 2016 | 05-10-2016 15:56 | 41.87781 | -87.6558 | (41.877811861, -87.655758012) |
| 3 | 05-03-2016 09:00 | THEFT | STREET | FALSE | 1 | 2016 | 05-10-2016 15:56 | 41.84302 | -87.6172 | (41.843016958, -87.61722727) |
| 4 | 05-03-2016 22:08 | THEFT | STREET | FALSE | 14 | 2016 | 05-10-2016 15:56 | 41.9109 | -87.686 | (41.910900826, -87.686018747) |
| 5 | 05-03-2016 21:45 | THEFT | STREET | FALSE | 14 | 2016 | 05-10-2016 15:56 | 41.90824 | -87.6784 | (41.908237096, -87.678437417) |
| 6 | 05-03-2016 18:30 | THEFT | RESIDENC | FALSE | 19 | 2016 | 05-10-2016 15:56 | 41.92732 | -87.6658 | (41.927321839, -87.665810418) |
| 7 | 05-03-2016 21:00 | THEFT | STREET | FALSE | 16 | 2016 | 05-10-2016 15:56 | 41.99596 | -87.7975 | (41.995961517, -87.797528563) |
| 8 | 05-03-2016 07:00 | THEFT | VEHICLE-C | FALSE | 22 | 2016 | 05-10-2016 15:56 | 41.72138 | -87.6498 | (41.721379795, -87.649781161) |
| 9 | 05-03-2016 20:30 | THEFT | PARKING L | FALSE | 19 | 2016 | 05-10-2016 15:56 | 41.92528 | -87.6594 | (41.925278177, -87.65936796) |
| 10 | 05-03-2016 19:30 | THEFT | STREET | FALSE | 25 | 2016 | 05-10-2016 15:56 | 41.90528 | -87.7326 | (41.90528413, -87.732615525) |
| 11 | 05-03-2016 21:45 | THEFT | RESIDENC | FALSE | 19 | 2016 | 05-10-2016 15:56 | 41.94999 | -87.6532 | (41.949986394, -87.65319658) |
| 12 | 05-03-2016 06:30 | THEFT | STREET | FALSE | 9 | 2016 | 05-10-2016 15:56 | 41.82034 | -87.6418 | (41.820343226, -87.641759214) |
| 13 | 05-03-2016 21:00 | THEFT | PARKING L | FALSE | 3 | 2016 | 05-10-2016 15:56 | 41.77107 | -87.5683 | (41.771073064, -87.568278663) |
| 14 | 05-03-2016 19:00 | THEFT | CHA PARKI | FALSE | 5 | 2016 | 05-10-2016 15:56 | 41.65325 | -87.6048 | (41.653249978, -87.604828177) |
| 15 | 05-03-2016 19:00 | THEFT | STREET | FALSE | 17 | 2016 | 05-10-2016 15:56 | 41.94646 | -87.7207 | (41.946461543, -87.72072354) |
| 16 | 05-03-2016 20:30 | THEFT | STREET | FALSE | 16 | 2016 | 05-10-2016 15:56 | 41.9937 | -87.7942 | (41.993703853, -87.794185238) |
| 17 | 05-03-2016 13:00 | THEFT | STREET | FALSE | 18 | 2016 | 05-10-2016 15:56 | 41.88945 | -87.628 | (41.889453169, -87.627994833) |
| 18 | 05-03-2016 22:30 | THEFT | RESIDENC | FALSE | 18 | 2016 | 05-10-2016 15:56 | 41.91995 | -87.6516 | (41.919949978, -87.651649056) |
| 19 | 05-04-2016 11:15 | THEFT | SMALL RE | TRUE | 6 | 2016 | 05-11-2016 15:50 | 41.74271 | -87.6341 | (41.742710224, -87.634088181) |
| 20 | 05-03-2016 19:30 | THEFT | CTA TRAIN | FALSE | 1 | 2016 | 05-10-2016 15:56 | 41.86817 | -87.6274 | (41.868165405, -87.62743954) |
| 21 | 05-03-2016 06:30 | THEFT | STREET | FALSE | 1 | 2016 | 05-10-2016 15:56 | 41.88316 | -87.6445 | (41.883164818, -87.644508395) |
| 22 | 05-03-2016 18:30 | THEFT | STREET | FALSE | 16 | 2016 | 05-10-2016 15:56 | 41.99057 | -87.7934 | (41.99057177, -87.793398734) |
| 23 | 05-03-2016 14:00 | THEFT | STREET | FALSE | 6 | 2016 | 05-10-2016 15:56 | 41.74771 | -87.6658 | (41.747712344, -87.665798083) |
| 24 | 05-03-2016 09:00 | THEFT | APARTMEI | FALSE | 6 | 2016 | 05-10-2016 15:56 | 41.75777 | -87.6078 | (41.757767253, -87.607818223) |
| 25 | 05-03-2016 00:30 | THEFT | CTA TRAIN | FALSE | 10 | 2016 | 05-10-2016 15:56 | 41.85443 | -87.6857 | (41.854427952, -87.685700213) |
| 26 | 05-03-2016 12:15 | THEFT | APARTMEI | FALSE | 7 | 2016 | 05-10-2016 15:56 | 41.78273 | -87.6497 | (41.782728365, -87.649720276) |

**FIG 4.1 Date Set**

### Data Preprocessing Module

8000 entries are present in the dataset. The null values are removed using df = df.dropna() where df is the data frame. The categorical attributes (Location, Block, Crime Type, Community Area) are converted into numeric using Label Encoder. The date

attribute is splitted into new attributes like month and hour which can be used as feature for the model.

**Feature selection Module**

Features selection is done which can be used to build the model. The attributes used for feature selection are Block, Location, District, Community area, X co-ordinate , Y coordinate, Latitude , Longitude, Hour and month.

**Building and Training Model**

After feature selection location and month attribute are used for training. The dataset is divided into pair of xtrain ,ytrain and xtest, y test. The algorithms model is imported form skleran. Building model is done using model. Fit (xtrain, ytrain).

**Prediction Module**

After the model is build using the above process, prediction is done using model.predict(xtest). The accuracy is calculated using accuracy_score imported from metrics - metrics.accuracy_score (ytest, predicted).

**Visualization Module**

Using matpoltlib library from sklearn. Analysis of the crime dataset is done by plotting various graphs.

## 4.2 SYSTEM STUDY

### Feasibility Study

The feasibility of the project is analyzed in this phase and business proposal is put forth with a very general plan for the project and some cost estimates. During system analysis the feasibility study of the proposed system is to be carried out. This is to ensure that the proposed system is not a burden to the company. For feasibility analysis, some understanding of the major requirements for the system is essential.

Three key considerations involved in the feasibility analysis are

- Economical Feasibility
- Technical Feasibility
- Social Feasibility

**Economical Feasibility**

This study is carried out to check the economic impact that the system will have on the organization. The amount of fund that the company can pour into the research and development of the system is limited. The expenditures must be justified. Thus the developed system as well within the budget and this was achieved because most of the technologies used are freely available. Only the customized products had to be purchased.

**Technical Feasibility**

This study is carried out to check the technical feasibility, that is, the technical requirements of the system. Any system developed must not have a high demand on the available technical resources. This will lead to high demands on the available technical resources. This will lead to high demands being placed on the client. The developed system must have a modest requirement, as only minimal or null changes are required for implementing this system.

**Social Feasibility**

The aspect of study is to check the level of acceptance of the system by the user. This includes the process of training the user to use the system efficiently. The user must not feel threatened by the system, instead must accept it as a necessity. The level of acceptance by the users solely depends on the methods that are employed to educate the user about the system and to make him familiar with it. His level of confidence must be raised so that he is also able to make some constructive criticism, which is welcomed, as he is the final user of the system.

# CHAPTER 5

# SOFTWARE TESTING

The test scenario is a detailed document of test cases that cover end to end functionality of a software application in liner statements. The liner statement is considered as a scenario. The test scenario is a high-level classification of testable requirements. These requirements are grouped on the basis of the functionality of a module and obtained from the use cases. In the test scenario, there is a detailed testing process due to many associated test cases. Before performing the test scenario, the tester has to consider the test cases for each scenario.

Documentation testing can start at the very beginning of the software process and hence save large amounts of money, since the earlier a defect is found the less it will cost to be fixed. The most popular testing documentation files are test reports, plans, and checklists. These documents are used to outline the team's workload and keep track of the process. Let's take a look at the key requirements for these files and see how they contribute to the process. Test strategy. An outline of the full approach to product testing. As the project moves along, developers, designers, product owners can come back to the document and see if the actual performance corresponds to the planned activities.

Test data. The data that testers enter into the software to verify certain features and their outputs. Examples of such data can be fake user profiles, statistics, media content, similar to files that would be uploaded by an end-user in a ready solution.

Test plans. A file that describes the strategy, resources, environment, limitations, and schedule of the testing process. It's the fullest testing document, essential for informed planning. Such a document is distributed between team members and shared with all stakeholders.

Test scenarios. In scenarios, testers break down the product's functionality and interface by modules and provide real-time status updates at all testing stages. A module can be described by a single statement, or require hundreds of statuses, depending on its size and scope.

- Testing can be done in the early phases of the software development lifecycle when other modules may not be available for integration
- Fixing an issue in Unit Testing can fix many other issues occurring in later development and testing stages

the system or acceptance testing

- Code coverage can be measured
- Fewer bugs in the System and Acceptance testing
- Code completeness can be demonstrated using unit tests. This is more useful in the agile process. Testers don't get the functional builds to test until integration is completed.
- Code completion cannot be justified by showing that you have written and checked in the code. But running Unit tests can demonstrate code completeness.
- Expect robust design and development as developers write test cases by understanding the specifications first.
- Easily identify who broke the build
- Saves development time: Code completion may take more time but due to decreased defect count overall development time can be saved.

**5.1 GENERAL**

Unit Testing frameworks are mostly used to help write unit tests quickly and easily. Most of the programming languages do not support unit testing with the inbuilt compiler. Third-party open source and commercial tools can be used to make unit testing even more fun.

List of popular Unit Testing tools for different programming languages:

- Java framework – JUnit
- PHP framework – PHPUnit
- C++ frameworks – UnitTest++ and Google C++
- .NET framework – NUnit
- Python framework – py.test

Functional Testing is a type of black box testing whereby each part of the system is tested against functional specification/requirements. For instance, seek answers to the following questions,

Are you able to login to a system after entering correct credentials?
Does your payment gateway prompt an error message when you enter incorrect card number? Does your “add a customer” screen adds a customer to your records

Well, the above questions are mere samples to perform full-fledged functional testing of a system.

## 5.2 TEST DRIVEN DEVELOPMENT

Test Driven Development, or TDD, is a code design technique where the programmer writes a test before any production code, and then writes the code that will make that test pass. The idea is that with a tiny bit of assurance from that initial test, the programmer can feel free to refactor and refactor some more to get the cleanest code they know how to write. The idea is simple, but like most simple things, the execution is hard. TDD requires a completely different mind set from what most people are used to and the tenacity to deal with a learning curve that may slow you down at first.

## 5.3 UNIT TESTING

Unit testing is a level of software testing where individual units/ components of a software are tested. The purpose is to validate that each unit of the software performs as designed. A unit is the smallest testable part of any software. It usually has one or a few inputs and usually a single output. In procedural programming, a unit may be an individual program, function, procedure, etc. In object-oriented programming, the smallest unit is a method, which may belong to a base/ super class, abstract class or derived/ child class. (Some treat a module of an application as a unit. This is to be discouraged as there will probably be many individual units within that module.) Unit testing frameworks, drivers, stubs, and mock/ fake objects are used to assist in unit testing.

A unit can be almost anything you want it to be -- a line of code, a method, or a class. Generally though, smaller is better. Smaller tests give you a much more granular view of how your code is performing. There is also the practical aspect that when you test very small units, your tests can be run fast; like a thousand tests in a second fast.

Black Box testers don't care about Unit Testing. Their main goal is to validate the application against the requirements without going into the implementation details. Unit Testing is not a new concept. It's been there since the early days of programming. Usually, developers and sometimes White box testers write Unit tests to improve code quality by verifying each and every unit of the code used to implement functional requirements (aka test drove development TDD or test-first development). Most of us might know the classic definition of Unit Testing – “Unit Testing is the method of verifying the smallest piece of testable code against its purpose.” If the purpose or

In simple words, Unit Testing means – writing a piece of code (unit test) to verify the code (unit) written for implementing requirements.

**5.4 BLACKBOX TESTING:**

During functional testing, testers verify the app features against the user specifications. This is completely different from testing done by developers which is unit testing. It checks whether the code works as expected. Because unit testing focuses on the internal structure of the code, it is called the white box testing. On the other hand, functional testing checks app's functionalities without looking at the internal structure of the code, hence it is called black box testing. Despite how flawless the various individual code components may be, it is essential to check that the app is functioning as expected, when all components are combined. Here you can find a detailed comparison between functional testing vs unit testing.

**5.5 INTEGRATION TESTING:**

Integration Testing is a level of software testing where individual units are combined and tested as a group. The purpose of this level of testing is to expose faults in the interaction between integrated units. Test drivers and test stubs are used to assist in Integration Testing. Testing performed to expose defects in the interfaces and in the interactions between integrated components or systems. See also component integration testing, system integration testing.

**COMPONENT INTEGRATION TESTING:**

Testing performed to expose defects in the interfaces and interaction between integrated components. System integration testing: Testing the integration of systems and packages; testing interfaces to external organizations (e.g. Electronic Data Interchange, Internet).

Integration tests determine if independently developed units of software work correctly when they are connected to each other. The term has become blurred even by the diffuse standards of the software industry, so I've been wary of using it in my writing. In particular, many people assume integration tests are necessarily broad in scope, while they can be more effectively done with a narrower scope.

As often with these things, it's best to start with a bit of history. When I first learned about integration testing, it was in the 1980's and the waterfall was the dominant influence of software development thinking. In a larger project, we would have a design phase that would specify the interface and behavior of the various modules in

unusual for one programmer to be responsible for a single module, but this would be big enough that it could take months to build it. All this work was done in isolation, and when the programmer believed it was finished they would hand it over to QA for testing.

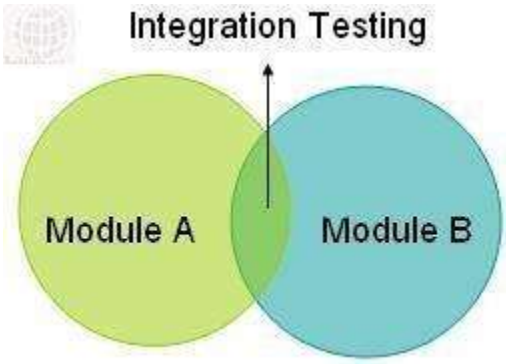



**Fig 5.1 Integration Testing**

**5.6 SYSTEM TESTING**

System Testing is a level of software testing where a complete and integrated software is tested. The purpose of this test is to evaluate the system's compliance with the specified requirements. System Testing means testing the system as a whole. All the modules/components are integrated in order to verify if the system works as expected or not. System Testing is done after Integration Testing. This plays an important role in delivering a high-quality product.

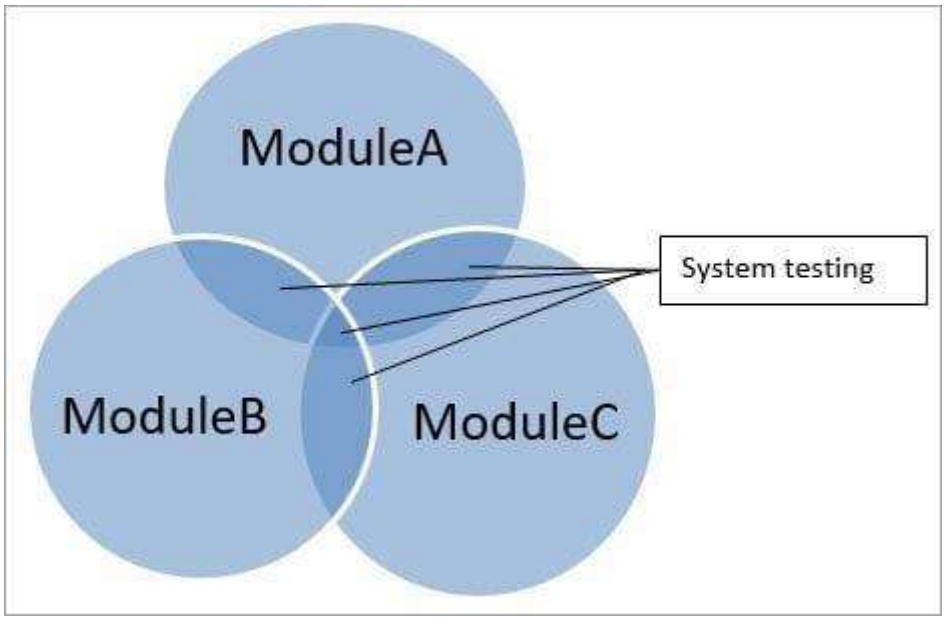



**Fig 5.2 Module Testing**

System testing is a method of monitoring and assessing the behaviour of the complete and fully-integrated software product or system, on the basis of pre-decided specifications and functional requirements. It is a solution to the question "whether the complete system functions in accordance to its pre-defined requirements?"

It's comes under black box testing i.e. only external working features of the software are evaluated during this testing. It does not requires any internal knowledge of the coding, programming, design, etc., and is completely based on users-perspective.

A black box testing type, system testing is the first testing technique that carries out the task of testing a software product as a whole. This System testing tests the integrated system and validates whether it meets the specified requirements of the client.

System testing is a process of testing the entire system that is fully functional, in order to ensure the system is bound to all the requirements provided by the client in the form of the functional specification or system specification documentation. In most cases, it is done next to the Integration testing, as this testing should be covering the end-to-end system's actual routine. This type of testing requires a dedicated Test Plan and other test documentation derived from the system specification document that should cover both software and hardware requirements. By this test, we uncover the errors. It ensures that all the system works as expected. We check System performance and functionality to get a quality product. System testing is nothing but testing the system as a whole. This testing checks complete end-to-end scenario as per the customer's point of view. Functional and Non-Functional tests also done by System testing. All things are done to maintain trust within the development that the system is defect-free and bug-free. System testing is also intended to test hardware/software requirements specifications. System testing is more of a limited type of testing; it seeks to detect both defects within the "inter-assemblages".

## 5.7 REGRESSION TESTING

Regression Testing is a type of testing that is done to verify that a code change in the software does not impact the existing functionality of the product. This is to make sure the product works fine with new functionality, bug fixes or any change in the existing feature. Previously executed test cases are re-executed in order to verify the impact of change.

Regression Testing is a Software Testing type in which test cases are re-executed in order to check whether the previous functionality of the application is working fine and the new changes have not introduced any new bugs. This test can be performed on a new build when there is a significant change in the original functionality that too even in a single bug fix. For regression testing to be effective, it needs to be seen as one part of a comprehensive testing methodology that is cost-effective and efficient while still incorporating enough variety—such as well-designed frontend UI automated tests alongside targeted unit testing, based on smart risk prioritization—to prevent any aspects of your software applications from going unchecked. These days, many Agile work environments employing workflow practices such as XP (Extreme Programming), RUP (Rational Unified Process), or Scrum appreciate regression testing as an essential aspect of a dynamic, iterative development and deployment schedule.

But no matter what software development and quality-assurance process your organization uses, if you take the time to put in enough careful planning up front, crafting a clear and diverse testing strategy with automated regression testing at its core, you can help prevent projects from going over budget, keep your team on track, and, most importantly, prevent unexpected bugs from damaging your products and your company's bottom line. Performance testing is the practice of evaluating how a system performs in terms of responsiveness and stability under a particular workload. Performance tests are typically executed to examine speed, robustness, reliability, and application size. The process incorporates "performance" indicators such as:

Load Testing is type of performance testing to check system with constantly increasing the load on the system until the time load is reaches to its threshold value. Here Increasing load means increasing number of concurrent users, transactions & check the behavior of application under test. It is normally carried out underneath controlled environment in order to distinguish between two different systems. It is also called as "Endurance testing" and "Volume testing". The main purpose of load testing is to monitor the response time and staying power of application when system is performing well under heavy load. Load testing comes under the Non Functional Testing & it is designed to test the non-functional requirements of a software application.

Load testing is perform to make sure that what amount of load can be withstand the application under test. The successfully executed load testing is only if the specified test cases are executed without any error in allocated time.

# CHAPTER 6

# CONCLUSION AND FUTURE WORK

## 6.1 Conclusion

In this paper focused on building predictive models for crime frequencies per crime type per month. The crime rates in India are increasing day by day due to many factors such as increase in poverty, implementation, corruption, etc. The proposed model is very useful for both the investigating agencies and the police official in taking necessary steps to reduce crime. The project helps the crime analysis to analysis these crime networks by means of various interactive visualization. Future enhancement of this research work on training bots to predict the crime prone areas by using machine learning techniques. Since, machine learning is similar to data mining advanced concept of machine learning can be used for better prediction. The data privacy, reliability, accuracy can be improved for enhanced prediction.

## 6.2 Future Enhancement

Crime analysis takes past crime data to predict future crime locations and time. Crime prediction for future crime is process that finds out crime rate change from one year to the next and projects those changes into the future. Crime predictions can be made through both qualitative and quantitative methods.

## REFERENCES

[1] Ginger Saltos and Mihaela Coacea, An Exploration of Crime prediction Using Data Mining on Open Data, International journal of Information technology & Decision Making,2017.

[2] Shiju Sathyadevan, Devan M.S, Surya Gangadharan.S, Crime Analysis and Prediction Using Data Mining, First International Conference on networks & soft computing (IEEE) 2014.

[3] Khushabu A.Bokde, Tisksha P.Kakade, Dnyaneshwari S. Tumasare, Chetan G.Wadhai B.E Student, Crime Detection Techniques Using Data Mining and K-Means, International Journal of Engineering Research & technology (IJERT) ,2018.

[4] H.Benjamin Fredrick David and A.Suruliandi,Survey on crime analysis and prediction using data mining techniques, ICTACT Journal on Soft computing, 2017.

[5] Tushar Sonawanev, Shirin Shaikh, rahul Shinde, Asif Sayyad, Crime Pattern Analysis, Visualization And prediction Using Data Mining, Indian Journal of Computer Science and Engineering (IJCSE), 2015.

[6] RajKumar.S, Sakkarai Pandi.M, Crime Analysis and prediction using data mining techniques, International Journal of recent trends in engineering & research,2019.

[7] Sarpreet kaur, Dr. Williamjeet Singh, Systematic review of crime data mining, International Journal of Advanced Research in computer science , 2015.

[8] Ayisheshim Almaw, Kalyani Kadam, Survey Paper on Crime Prediction using Ensemble Approach, International journal of Pure and Applied Mathematics,2018.

# APPENDIX

## A.SCREEN SHOTS

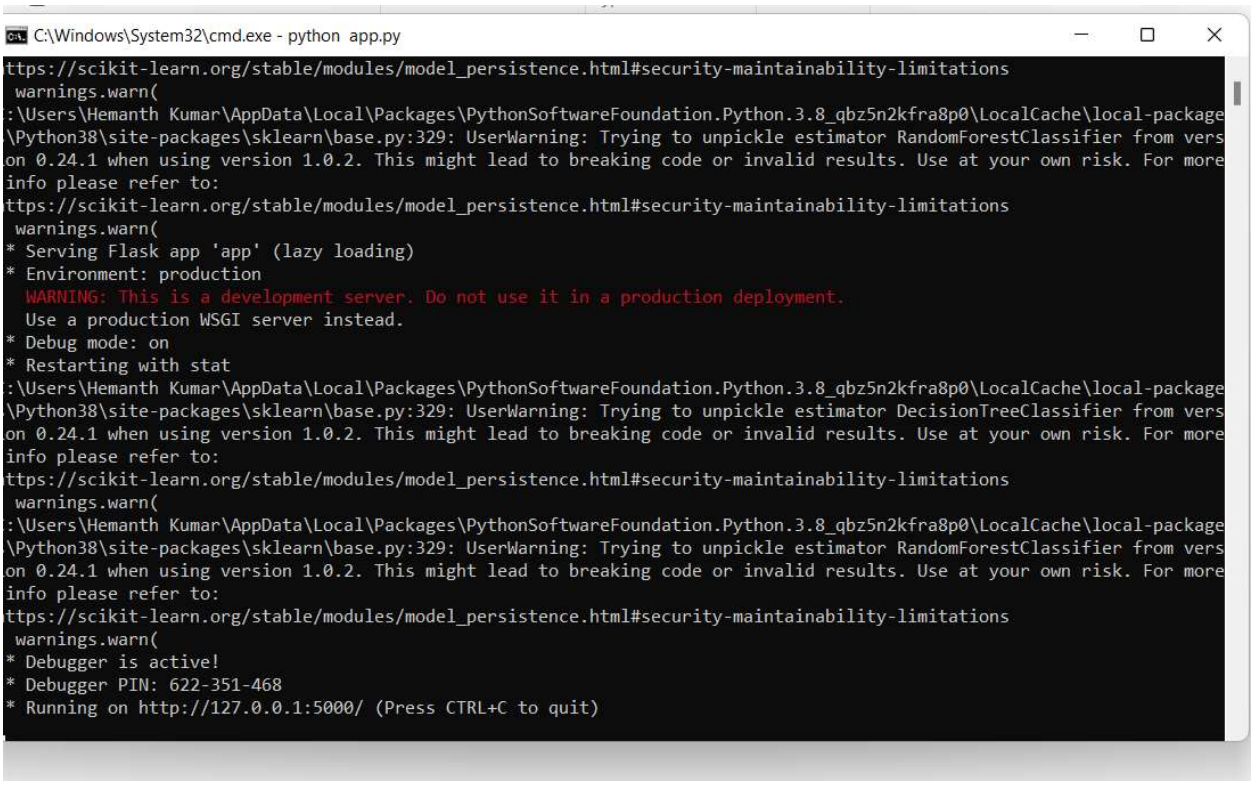



**Figure-1 shows the page of running our application**

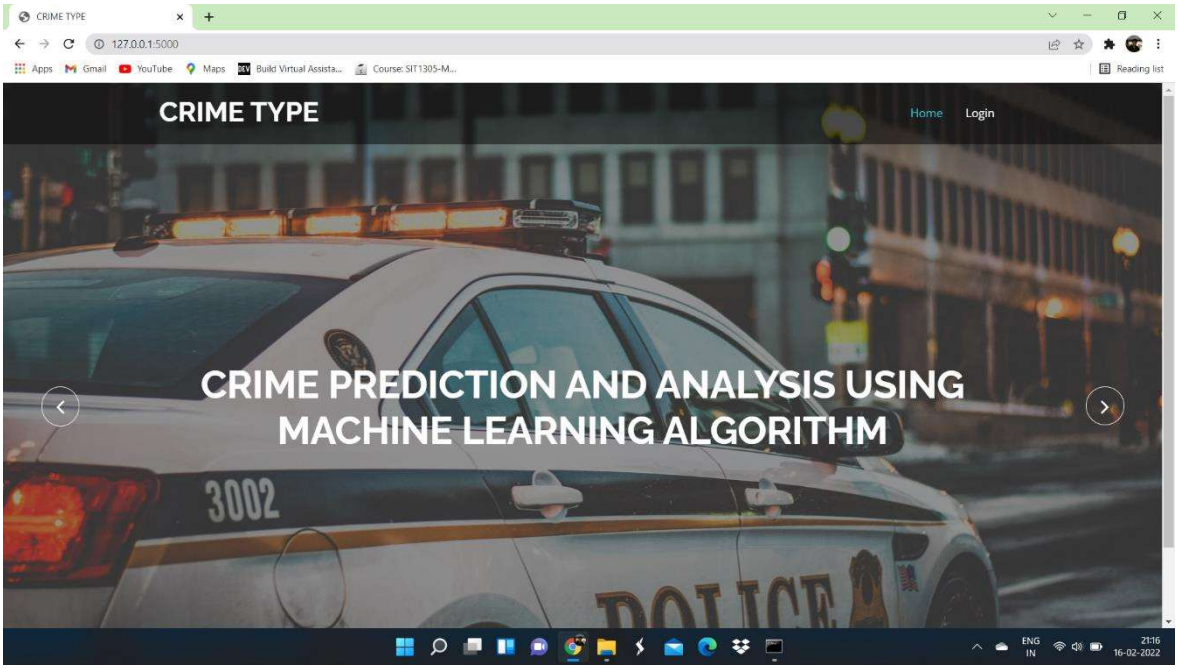



**Figure-2 shows the homepage of our application of crime analysis and prediction.**

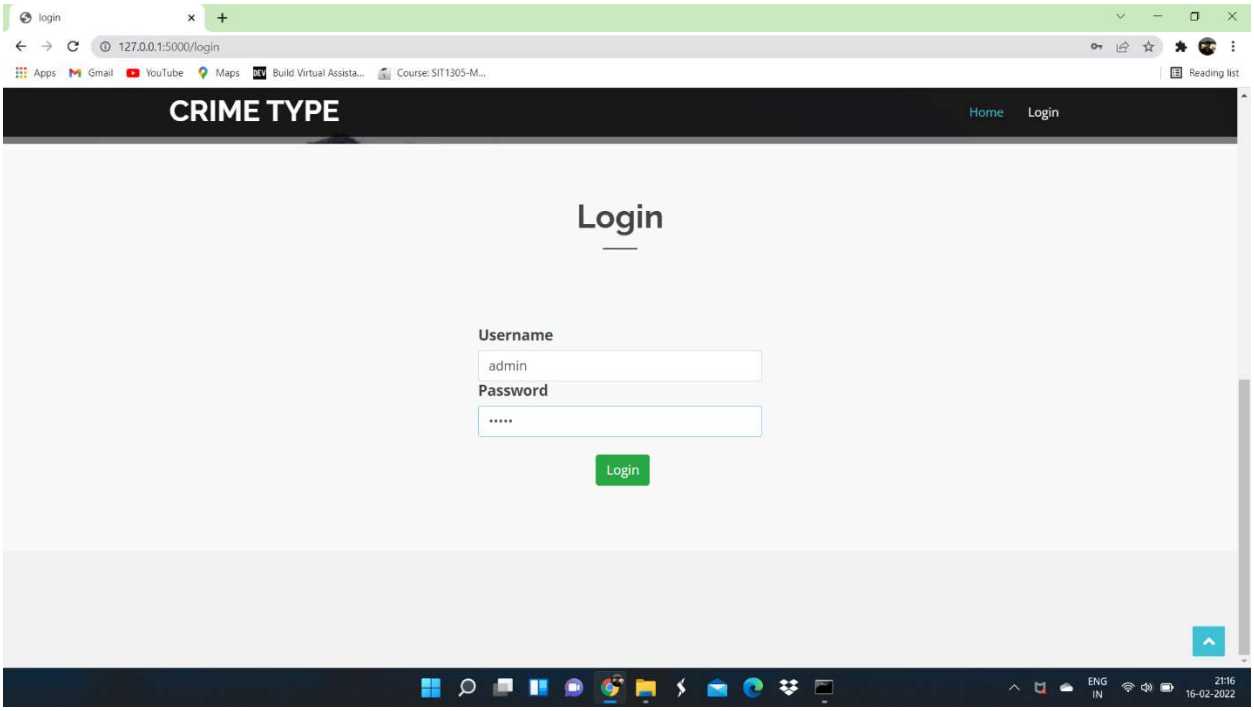



**Figure-3 shows the login page for predicting the crime.**

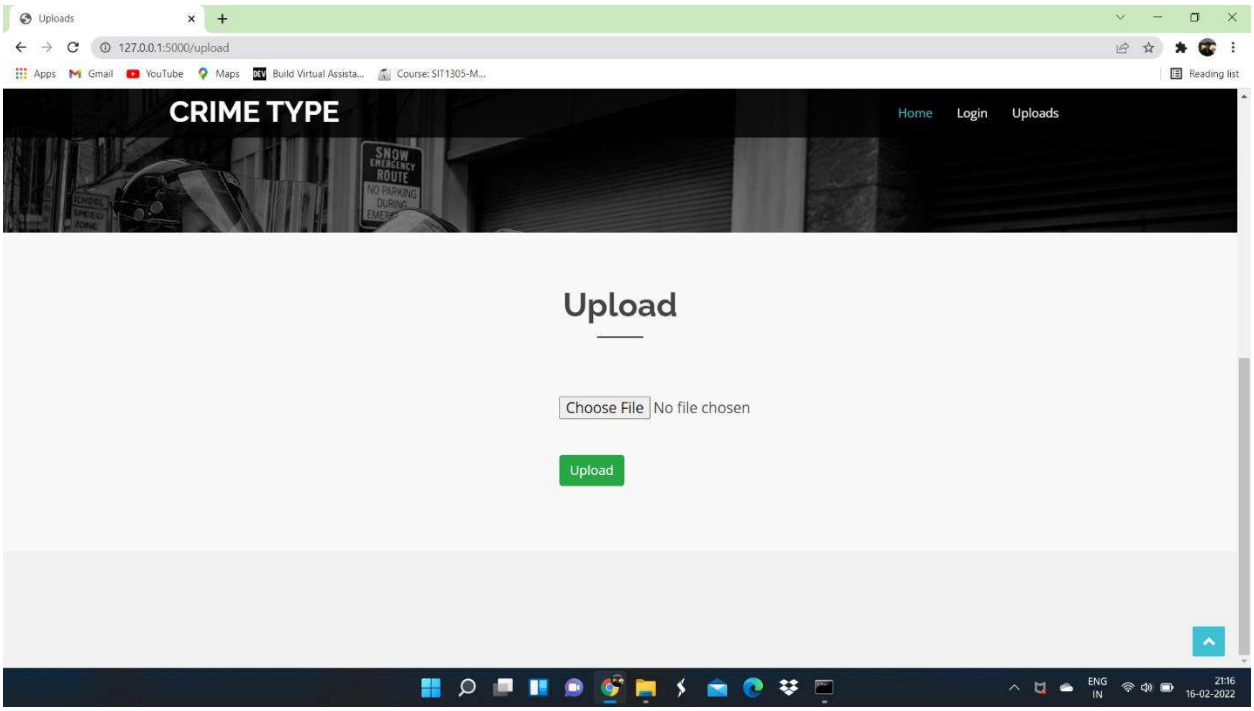
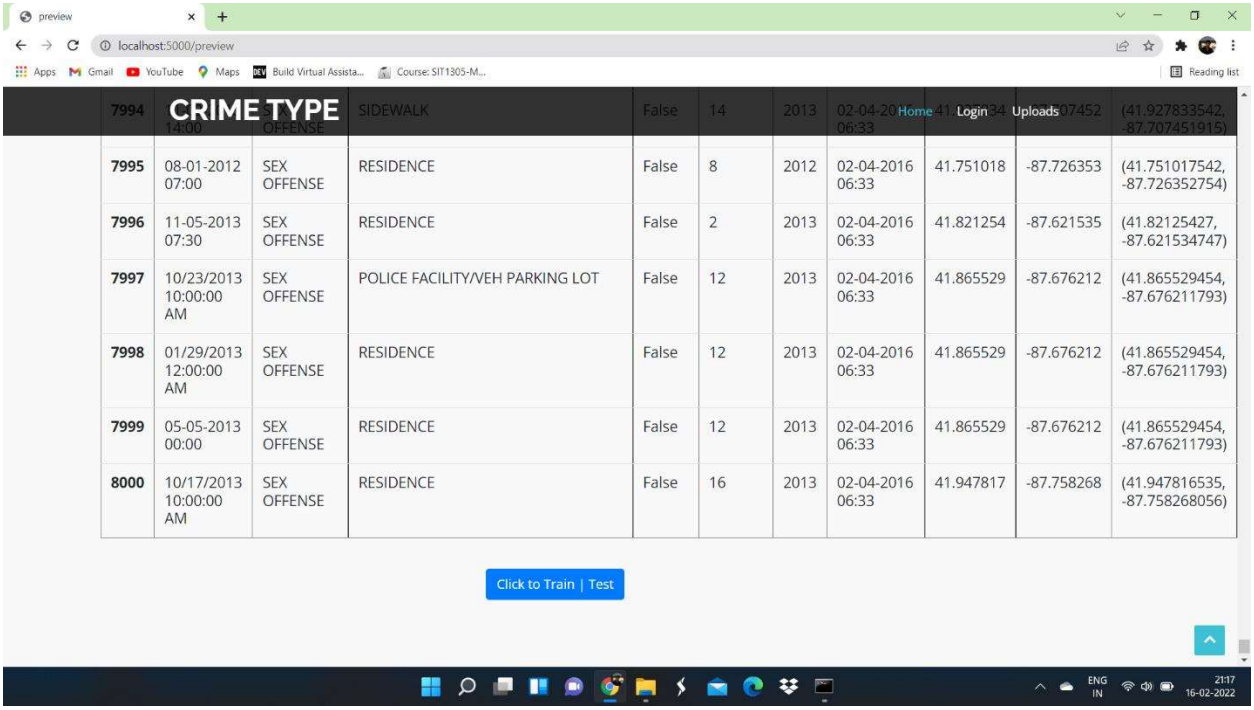



**Figure-4 shows the page to upload the dataset of crime prediction**




| | | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|
| 7994 | 14:00 | OFFENSE | SIDEWALK | False | 14 | 2013 | 02-04-20 06:33 | 41 | 34 | 07452 (41.927833542, -87.707451915) |
| **7995** | 08-01-2012 07:00 | SEX OFFENSE | RESIDENCE | False | 8 | 2012 | 02-04-2016 06:33 | 41.751018 | -87.726353 | (41.751017542, -87.726352754) |
| **7996** | 11-05-2013 07:30 | SEX OFFENSE | RESIDENCE | False | 2 | 2013 | 02-04-2016 06:33 | 41.821254 | -87.621535 | (41.82125427, -87.621534747) |
| **7997** | 10/23/2013 10:00:00 AM | SEX OFFENSE | POLICE FACILITY/VEH PARKING LOT | False | 12 | 2013 | 02-04-2016 06:33 | 41.865529 | -87.676212 | (41.865529454, -87.676211793) |
| **7998** | 01/29/2013 12:00:00 AM | SEX OFFENSE | RESIDENCE | False | 12 | 2013 | 02-04-2016 06:33 | 41.865529 | -87.676212 | (41.865529454, -87.676211793) |
| **7999** | 05-05-2013 00:00 | SEX OFFENSE | RESIDENCE | False | 12 | 2013 | 02-04-2016 06:33 | 41.865529 | -87.676212 | (41.865529454, -87.676211793) |
| **8000** | 10/17/2013 10:00:00 AM | SEX OFFENSE | RESIDENCE | False | 16 | 2013 | 02-04-2016 06:33 | 41.947817 | -87.758268 | (41.947816535, -87.758268056) |

**Figure-5 shows the dataset to train and test using machine learning algorithms.**

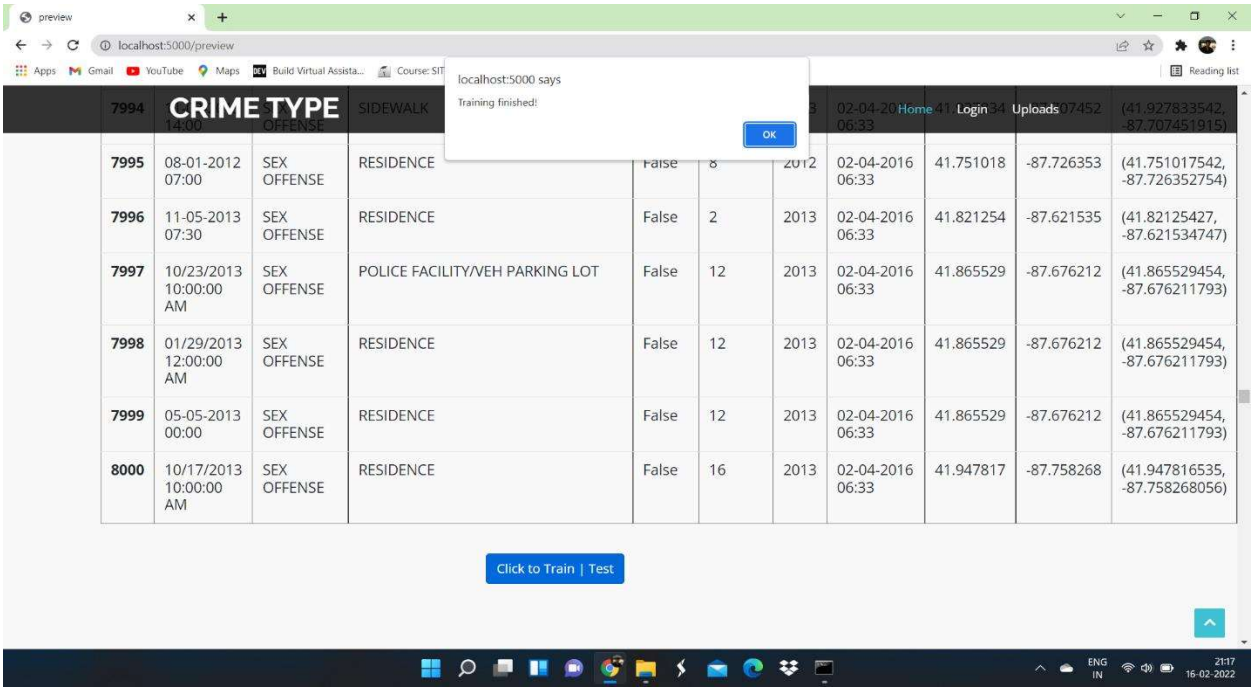



**Figure-6 shows the message to indicate the successful training using machine learning algorithms.**

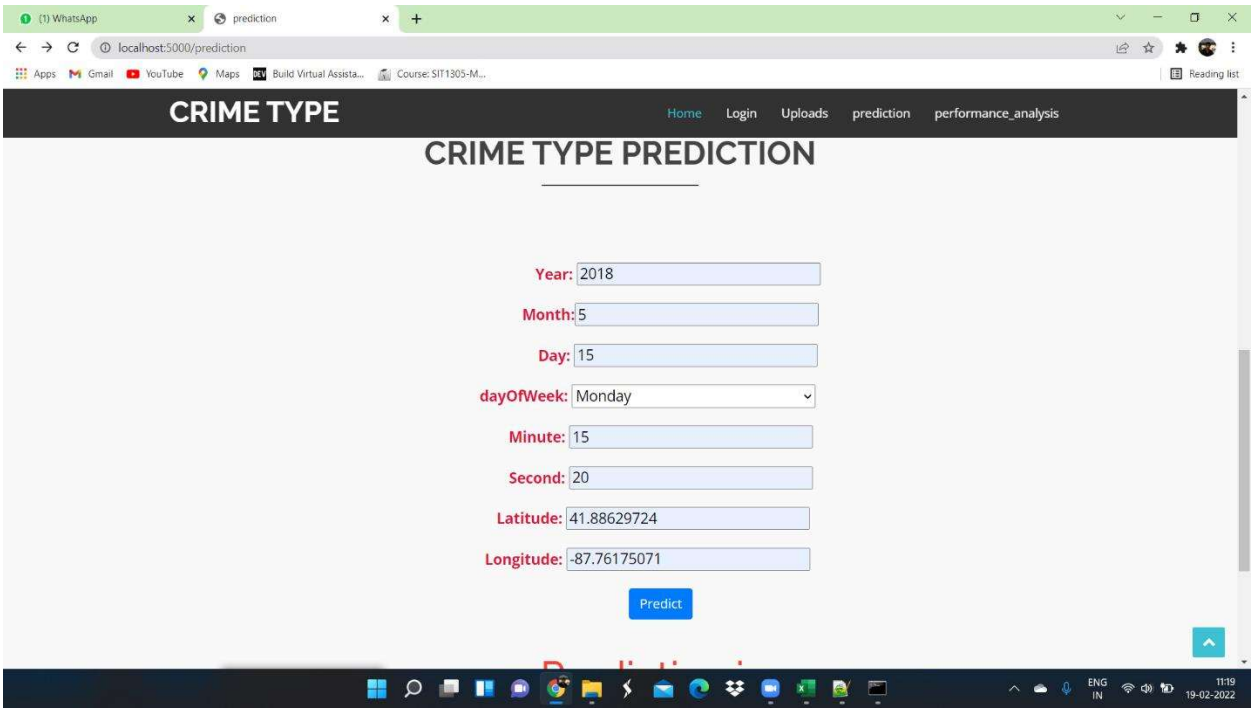



**Figure-7 shows the web page of filling the details about crime location, date and time to predict the crime.**

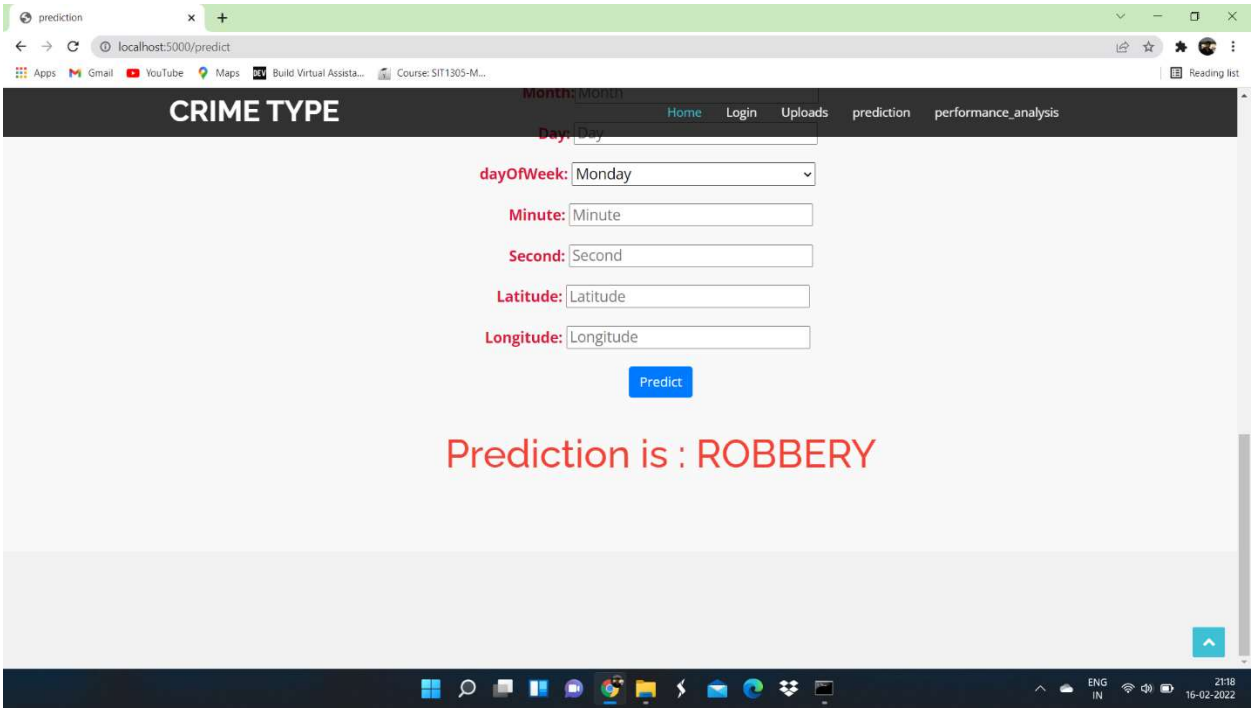



**Figure-8 shows the prediction of type of crime.**

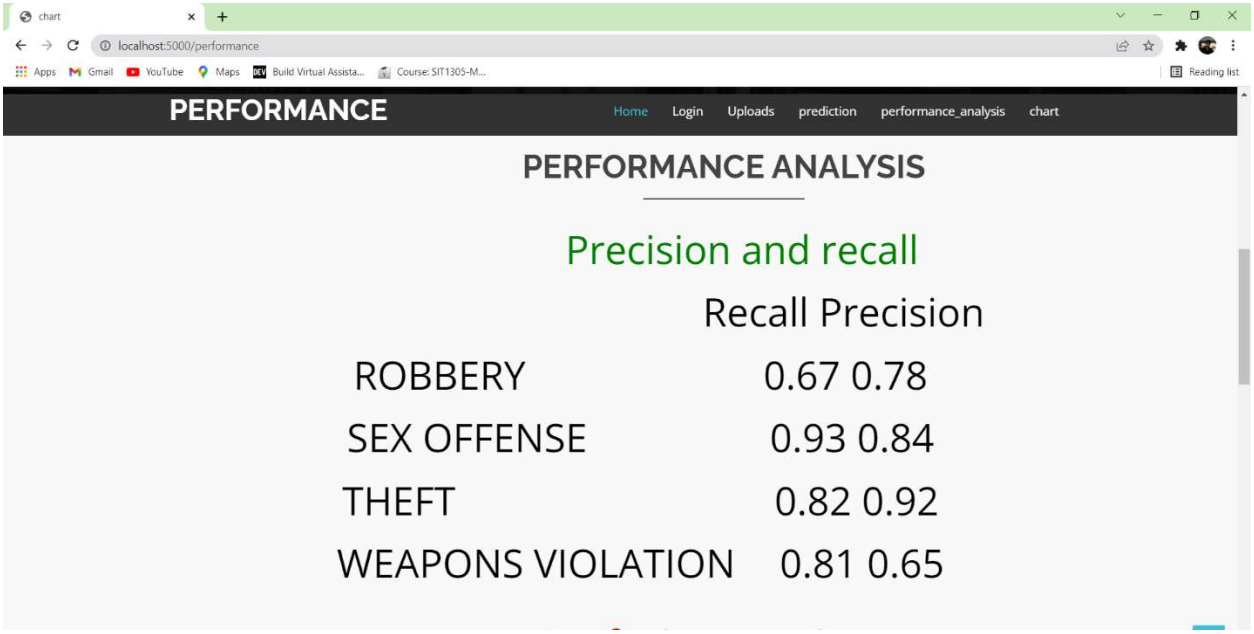



**Figure-9 shows the performance analysis for the type of crime**

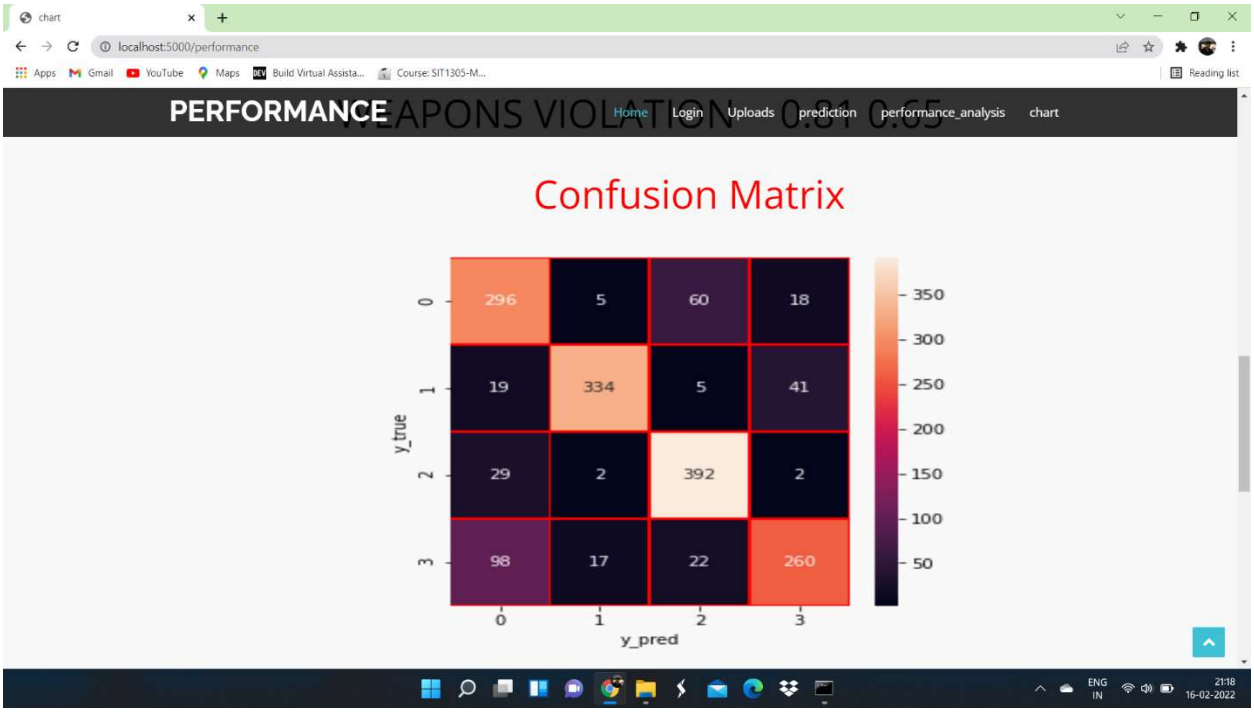



**Figure-10 shows the confusion matrix for the analysis of crime.**

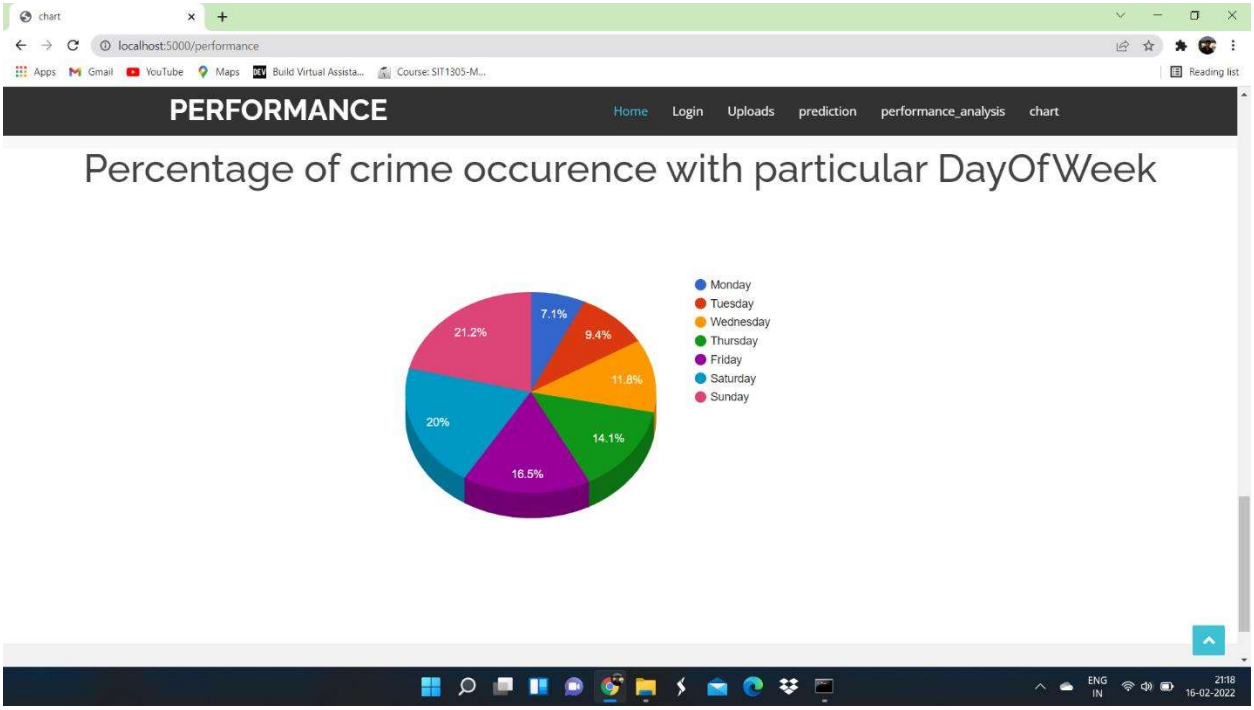



**Figure-11 shows the performance of crime occurrences with particular day of week.**

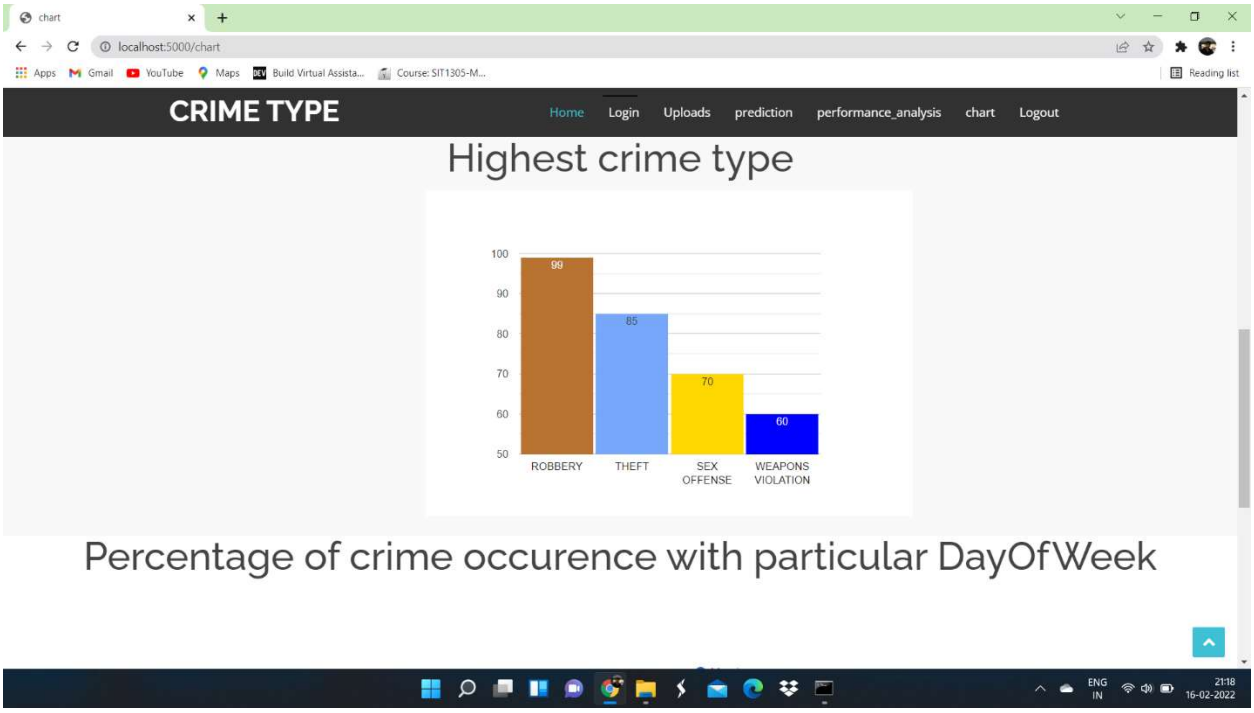



**Figure-12 shows the graph for percentage of crime occurrences with particular day of week.**

**B.SAMPLE CODE**

```
import numpy as np
import pandas as pd
from flask import Flask, request, jsonify, render_template, redirect, flash, send_file
from sklearn.preprocessing import MinMaxScaler
from sklearn.naive_bayes import GaussianNB
import pickle
app = Flask(_name_) #Initialize the flask App

model = pickle.load( open('model.pkl', 'rb') )
@app.route('/')

@app.route('/first')
def first():
        return render_template('first.html')



@app.route('/login')
def login():
        return render_template('login.html')
@app.route('/upload')
def upload():
   return render_template('upload.html')
@app.route('/preview',methods=["POST"])
def preview():
```

```
    if request.method == 'POST':
        dataset = request.files['datasetfile']
        df = pd.read_csv(dataset,encoding = 'unicode_escape')
        df.set_index('Id', inplace=True)
        return render_template("preview.html",df_view = df)



#@app.route('/home')
#def home():
 #   return render_template('home.html')

@app.route('/prediction', methods = ['GET', 'POST'])
def prediction():
    return render_template('prediction.html')



#@app.route('/upload')
#def upload_file():
#   return render_template('BatchPredict.html')



@app.route('/predict',methods=['POST'])
def predict():
            int_feature = [x for x in request.form.values()]
            print(int_feature)
```

```
        int_feature = [float(i) for i in int_feature]
        final_features = [np.array(int_feature)]
        prediction = model.predict(final_features)

        output = format(prediction[0])
        print(output)
        return render_template('prediction.html', prediction_text= output)
@app.route('/chart')
def chart():
        return render_template('chart.html')
@app.route('/performance')
def performance():
        return render_template('performance.html')

if _name_ == "_main_":
    app.run(debug=True)
```

## C.PLAGARISM REPORT:

# CRIME PREDICTION AND ANALYSIS USING MACHINE LEARNING

**Student name**

**Guide name**

## ABSTRACT

Crime examination and prediction is a strong method for knowing crime. The framework can anticipate spaces of high danger and comprehend spaces of high danger. Wrongdoing influences public prosperity, monetary development, and exposure. To shield society from wrongdoing, there is a requirement for better techniques and new strategies for further developing wrongdoing examination to ensure people in general. We give a framework that can examine, recognize and anticipate different wrongdoings in a specific district.

## Keywords

Crime prediction, Data mining, Open data, regression, decision tress, instance based learning, Naïve bayes

## INTRODUCTION

The quantity of violations is expanding step by step since current innovation and trend setting innovation assist crooks with submitting unlawful demonstrations. Rap, and so on Data about wrongdoing will be gathered on different online journals, news and sites. Huge information is utilized as a wrongdoing record. The information acquired with regards to data mining innovation will serve to rapidly recognize crooks and decrease wrongdoing episodes in wrongdoing inclined regions. Information digging is useful for the police division. One of the qualities of the police is the foundation of "problem areas" in regions where wrongdoing is probably going to happen. The utilization of data recovery techniques can accomplish critical outcomes in the wrongdoing revealing data set. The genuine advance in learning crime is to examine it. Wrongdoing investigation is the most common way of distinguishing, connecting, and recognizing the connections between various violations and qualities f. Investigation plans measurements, questions, and guides. It likewise assists with deciding if a wrongdoing should be perpetrated with a certain goal in mind or by another means. crime can be forecasted similarly that an abhorrent individual does it in a decent spot. Later triumph, they attempt to mirror crime in comparable circumstances. The presence of a wrongdoing relies upon many variables, like the culprit's insight and the security of the spot. This work follows the means of information investigation that fall into fundamental classifications, for example, information assortment, information arranging, information design, forecast, and portrayal. .

At the proposed level, different techniques for video observation and different strategies for wrongdoing forecast are utilized, utilizing AI calculations. Increments to our calculations are time

(hour, day, month, year), place (length, length), and wrongdoing arrangement.

Act 379 - Robbery
Act 13 – Gambling
Act 279 - Accident
Act 323 - Violence
Act 302 - Murder
Act 363 - Kidnapping

What happened is a classification that might have occurred. We are exploring different avenues regarding many phases of the calculation, like KNN (Nearest Neighbors), Certificate Tree, and standard woodland. We additionally do a great deal of work classes - first forecasting 6 sorts of potential offenses, and afterward attempting to separate among criminal and criminal offenses.

## LITERATURE SURVEY

**An Explor ation of Crime Prediction Using Data Mining on Open Data**

Ginger Saltos and Minhaela Cocea

As of late, the police have enrolled wrongdoing, recognized the quantity of violations, and planned the quantity of violations in both neighbourhood states through map-based projects and projects. neighbourhood police. what's more different business sectors utilizing official data. In this article you will find out with regards to the kinds of violations and how to manage them. The review was led utilizing Portsmouth University's SCIAMA High Performance e Computer Cluster and Weka programming.

**Crime Analysis and Prediction Using Data Mining**

Shiju Sathyade van,Deva n M.S, Surya Gangadh aran

As per the Crime Bureau, wrongdoings, for example, robbery, pyromania and different violations have diminished while wrongdoings, for example, murder, sex, savagery, assault and different violations have expanded. In this page data is gathered on different destinations, for example, sites, websites, news locales, web-based media destinations, RSS channels and then some. This data is utilized as a criminal record. The five-venture wrongdoing investigation is Data Collection, Classification, Model Identification, Prophecy, and Visualization. This paper has attempted to decide the classifications and expectations as per various tests. Grouping n is done dependent on Bayes' hypothesis which showed that 90% is valid.

**Crime Detect ion Techni ques Using data Minin g and KMeans**

Khushab u A.Bokde, Tisksha P. Kakade, Dnyanes hwari S. Tumsare, Chetan G. Wadhai

This report centers around wrongdoing examination, information assortment, and joining of calculations with the portrayed K. One of the reasons for wrongdoing examination is to distinguish and recognize a wrongdoing dependent on data about the idea of the wrongdoing and the discipline. Conversation implies sharing data or things in bunch numbers. Here is a rundown of data that resembles a bunch. K-Definition is the least complex and most ordinarily utilized calculation for bunching calculations underway programming. In light of the genuine reply, criminal data mining has a brilliant future with the expansion of execution and knowledge examination.

**Survey on crime analysis and predict ion using data mining techniques.**

Benjamin Fredrick David. H and Suruliand i

Criminal science is a strategy used to decide the idea of a wrongdoing and to distinguish it. Utilizing this data expulsion calculation,

it will actually want to distinguish wrongdoings and assist with recognizing crooks all the more rapidly. At the point when crooks leave, they make an imprint that can be utilized to look for hoodlums. This strategy is utilized to recognize crooks dependent on proof and data given by the local area. One of the strategies for wrongdoing examination utilized in this article is the text, content, non-credit strategy, and technique for wrongdoing. Various breaks down have shown that the utilization of GA to alter estimations builds the degree of exactness.

**Crime Pattern Analys is, Visual ization s And Predict ion Using Data Mining**

Tushar Sonawan ev, Shirin Shaikh,S haista Shaikh, Rahul Shinde, Asif Sayyad.

Wrongdoing examination is the review, relationship, and ID of the connection between various violations and the idea of wrongdoing. The fundamental reason for this review is to foster a technique for separating logical data that can generally take care of mind boggling issues identified with various kinds of wrongdoing. From this perusing and composing study, we can affirm that the subtleties of the wrongdoing are expanding and spreading, even in zota bytes. The appropriate response comes as compromise unique in relation to the spot where the wrongdoing was carried out. Offenses can be arranged by age, bunch, spot of wrongdoing, or kind of wrongdoing.

**Crime Analysis And predict ion using Data Mining Techniques**

Rajkumar .S, Sakkarai Pandi.M, Soundary a Jagan.J,V arnikasre e.P

The primary commitment of this article is to give another methodology dependent on the achievement of top to bottom investigation of different classes of work, for example, picture acknowledgment, object acknowledgment, picture handling, and normalization. Different calculations are utilized to dissect and contrast wrongdoing data with decide the best wrongdoing forecast calculation.

**Systematic Review of Crime Data Mining**

Sapreet kaur,Dr. Williamj eet Singh

This article portrays the strategies utilized, the hindrances to be tackled, the techniques utilized, and the archive for the evacuation and examination of criminal data. The approach comprises of three phases: the principal stage is the data search research, the subsequent stage is the stage, and the third stage is to introduce a rundown of the examination from wrongdoing, which incorporates investigation and announcing. research. The consequences of this review can assist new clients with understanding the innovation and greatness of criminal data recovery.

**Survey paper on Crime Prediction using Ensemble Approach**

Ayishesh im Almaw,K alyani kadam

Various phases of the critical thinking calculation are required dependent on the necessities for recognizing the wrongdoing data. One procedure can be more important than various strategies for tackling an issue. Autonomous models to accomplish ideal execution to defeat the supposed novel models. Research is being improved information assortment and recovery strategies.

## EXISTING SYSTEM

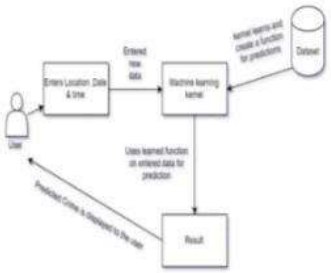
In criminal exploration, information mining can be classified into wrongdoing anticipation and wrongdoing counteraction. De Bruyn et al. al. A criminal framework was made utilizing another strategy for

looking at all individuals by their qualities and afterward joining them. Manish Gupta and others. al. The conventional framework utilized by the Indian police is characterized as an e-government program, and the utilization of the Internet is empowered dependent on issues, for example, criminal investigation apparatuses to help the police. He required the utilization of the Internet to acquire data from significant criminal documents under the sponsorship of the National Bureau of Investigation (NCDC) and to observe areas of interest utilizing criminal data extraction strategies like conglomeration. Sutapat Tiprungsri surveys the execution of bookkeeping investigation. model, particularly against control. The motivation behind his examination was to analyze whether deceitful channel innovation was utilized during the review. He utilized bunch investigation to assist evaluators with zeroing in on their work in regulating bunch medical coverage.

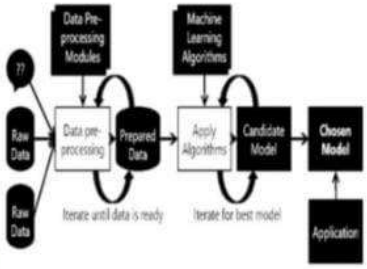
## PROPOSED SYSTEM

In this undertaking, we will recognize identified violations utilizing AI strategies. Data about the wrongdoing was gotten from the police. It comprises of criminal data like area, sort of wrongdoing, date, time, sum and length. Test data ought to be handled later these choices prior to handling, and estimations ought to be made to keep up with the respectability of the example. K-Near Neighborhoods (KNN) areas and different calculations (Certificate Tree and Random Memory) will be tried and utilized adequately in preparing to recognize guilty parties. The point of this undertaking is to furnish law requirement offices with a practical comprehension of how to utilize machine preparing, to rapidly distinguish and address violations, and to decrease the quantity of wrongdoings.

## SYSTEM ARCHITECTURE




## DATA FLOW DIAGRAM




2

## SYSTEM REQUIREMENTS

### HARDWARE REQUIREMENTS

System : Pentium Dual Core.

Hard Disk : 120 GB.

Monitor : 15'' LED

Input Devices : Keyboard, Mouse

Ram : 4 GB.

### SOFTWARE REQUIREMENTS

Operating system: Windows 7/10.

Coding Language: Python

## MODULES

Data Gathering Module

Data Preprocessing Module

Feature selection Module

Building and Training Model

Prediction Module

Visualization Module

## MODULE DESCRIPTIONS

**Data Gathering Module**

Kaggle's criminal data program is utilized as CSV.

**Data Preprocessing Module**

The informational collection contains 10,000 records. Erase the invalid worth utilizing df = df.dropna (), where df is the information range. The properties of the classifications (Location, Parking, Type of Crime, Common Territory) are determined utilizing the Label Encoder. The date property is partitioned into new things, for example, the month and time, which can be utilized as a format work.

**Feature selection Module**

Choice choices can be utilized to make a format. The capacities utilized in the choice are Parking, Location, Region, Association, X connection, Y interface, Size, Length, Time and month.

**Building and Training Module**

Properties are utilized in the preparation as indicated by the one-site choice method for the month. The informational collection is partitioned into two sections: ytrain, xtest and test. The calculation mode was presented in scleran mode. Configuration is finished utilizing a format. Appropriate (xtrain, ytrain).

**Prediction Module**

In the wake of making a model utilizing the above technique, make a forecast utilizing the model. Xtest. Truth is an amount - an item that is imported is estimated decently. Score_ focuses (test, anticipated).

**Visualization Module**

Sklearn's matplotlib library is utilized. Information base wrongdoing investigation is utilized to get ready different models.

## CONCLUSION

In this paper, we offer a practical way to deal with wrongdoing forecast by consolidating data from numerous areas and natural data. Our methodology includes recording crime in certain spaces and mirroring it dependent on top to bottom criminal schooling. Police watches can utilize criminal data to screen crime locations and further develop police watches. Our technique comprises of three stages. To begin with, we gathered a wide scope of data from the Chicago City Network, American FactFinder, Underground Air, and Google Street View. Picture data was utilized to remove ecological data. We then, at that point, examined the connection between the amassed wrongdoing utilizing insights. So we made an information base that can be utilized adequately to distinguish wrongdoing. DNN was then used to join various levels and loads to all the more likely incorporate underlying, spatial, and natural data to identify wrongdoings.

## RESULTS

Consequently, our full-time information preparing utilizing DNN 1: 2 proportion showed 84.25 precision, 74.35 exactness, 80.55 review, and 0.8333 AUC. All qualities transcend the qualities associated with the conventional technique. Likewise, we contrasted our DNN and a few models